DIAGNOSTIC AND PREDICTIVE MODELLING IN OSTEOARTHRITIS USING STATISTICAL AND MACHINE LEARNING TOOLS

PHILIPPA GRACE MCCABE

A thesis submitted in partial fulfilment of the requirements of Liverpool John Moores University for the degree of Doctor of Philosophy

December 2021

Acknowledgements

I would firstly like to thank my supervisors, Paulo, Ivan, and Bill for all of the help and advice that have made this work possible, and LJMU and EU H2020 for funding this work. I am beyond grateful for all of the guidance and advice that has been passed on over the past 3 years. If I've had any doubts, you have set me straight and helped me to believe that it will all work out. Ivan, thank you for the unplanned meetings where topics would span work, food, life and back to work whilst fixing all the issues I didn't know where there. To Ian, thank you for being there to provide pointers and helping me to hit the ground running with what the next chapter holds once this door of the PhD closes.

To my best friend, Zoé, thank you for being so understanding. Having you on the other end of a phone has helped me to keep going. Our walks and drives have put everything back in order and you have always been there even when I was a bit of a crank, so thank you. To be blessed with a best friend like you is an honour and I couldn't have got here without you. To Mark and Toby thank you for understanding and always having an extra seat in the library available when I needed somewhere else to work, keeping me laughing and keeping that coffee like rocket fuel and making sure that we are firmly in this together. To Kyle, thank you for being there. You do so much to keep me grounded and moving forward, so thank you for supporting and believing in me always.

My dear little Mama and Dad, thank you for all the support, especially over the past few years. You've been there to give me a boost when I've needed it, and escape when I want to bin everything and to celebrate all the little wins that have led us to this day. You both have been my cheerleaders in making me believe I can do it. Mum, you have been and always will be my inspiration, and especially with this work, thank you for motivating me on those days I really need it (the biscuits helped!). Dad, your humour has kept me going when I've needed to smile, and our little road trips have given me something to aim for when a break is a bit overdue, thank you. Thank you for making me believe I can do anything I set my mind and showing me that anything is possible. I will forever be grateful to you both for everything. I hope I never stop making you proud. I dedicate this thesis to you.

Without all of these people, and a few more, their support and encouragement, none of this would have been possible. So, to everyone, for everything, thank you.

'Don't dream it, be it!' **-** *Richard O'Brien | Tim Curry*

Abstract

Osteoarthritis (OA) is a degenerative bone disease that affects joints. OA is one of the most common diseases affecting people in old age. Between 12% and 30% of over 65s are affected by OA, with the knees being the most commonly affected joint. The process for making a diagnosis of knee osteoarthritis is time consuming and somewhat subjective. Clinicians assess a variety of clinical symptoms and information and establish if the patient meets the criteria for having the disease. The utilisation of a machine learning tool could potentially enhance the experience of patients in a clinical setting by reducing the amount of testing required to arrive at a firm diagnosis.

In clinical settings where patient education and behaviour modification are at the forefront, interpretable models are key, as it is vital to be able to explain a decision that leads to any course of action related to an individual patient. In Chapter 3, a model that could be used to aid clinicians in making a diagnosis is developed, and Chapter 4 a model to identify risk cohorts of people who do not yet have the disease is described. Chapter 5 takes those models and uses a different dataset to validate them and develop interactive web-based applications that have easy to explain results.

These models are expanded to consider the effect of gender in presentation of knee osteoarthritis and how this can influence the likelihood of presenting with the disease. Also, the use of multitask learning aims to describe the usefulness of combining datasets to enhance model performance.

Together, these models and approaches utilise both clinical and demographic features to help identify those with knee osteoarthritis and those who are at risk of developing the disease in a five-year timeframe. The models and apps have potential use in clinical settings both as a decision support tool and as a resource for patient education following UK validation of the model.

Publication and Dissemination of Results

Journal and Peer-reviewed Conference Articles

- **McCabe PG**, Lisboa P, Baltzopoulos V, Olier I.
  Externally validated models for first diagnosis and risk of progression of knee osteoarthritis.
  PLOS One. 2022. [Journal Paper]

- **McCabe PG**, Olier I, Ortega-Martorell S, Jarman I, Baltzopoulos V, Lisboa P.
  Comparative Analysis for Computer-Based Decision Support: Case Study of Knee Osteoarthritis.
  Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics). 2019. [Conference Article]

Talks and Conferences

- LJMU Faculty Research Day, Online, Liverpool, UK. May 2021. Oral Presentation. *Diagnostic And Predictive Modelling In Osteoarthritis Using Statistical And Machine Learning Tools*

- Open Research Week Research Café, Online, Liverpool, UK. February 2021. Oral Presentation. *OA: Open Access and Osteoarthritis*

- IDEAL 2019 Conference, November 2019. Oral Presentation. *Comparative Analysis for Computer-Based Decision Support: Case Study of Knee Osteoarthritis.*

Contents

Chapter 1: Background and introduction

1.1. Introduction

1.1.1. Disease and the burden

Osteoarthritis (OA) is a degenerative bone disease that affects joints as a whole. OA is one of the most common diseases affecting people in old age. The prevalence in people 65 years and older ranges from 12% to 30% [1]. The disease is also the most common form of arthritis to cause pain and mobility limitations. OA most commonly affects the knee, and around 10% of people over 55 years old have knee OA (KOA). This statistic is not surprising as weight-bearing joints, such as the knee or hip, are where disease occurs most [2]. The focus of this research is specifically KOA.

Weight is just one of the factors that can play a part in developing KOA. Some of the other factors are genetics, past injury and overuse of a specific joint [2]. Many of the risk factors of OA in any joint are non-modifiable, such as gender and a person's predisposition to other types of arthritis, for example rheumatoid. Despite many people thinking that OA is a disease that only affects the elderly, everyone is susceptible, with younger people more likely to develop the disease as a direct result of trauma. This type of OA is known as secondary OA. Most factors that cause OA are not features the patient can modify however there are some that, if dealt with, can slow the progression of OA, one such factor that can be changed is weight [2].

There are five stages of KOA according to the Kellgren-Lawrence (KL) scale [3]. These are differentiated between with the use of x-rays to determine the severity of the OA. Stage 0 is classed as no OA and Stage 4 is severe OA present in the joint. A visual example of how bones change due to OA is in Figure 1-1. A clinician usually analyses and classifies images for diagnosis. By using both humans and computer based models there is the potential for more reliable diagnoses [4]. Stage 1 is the point at which disease changes are likely to begin but go unnoticed as they do not typically cause symptoms to the patient. Stage 2 is usually the point of diagnosis as this is where symptoms usually begin to bother the patient. The advice usually given at this stage of the disease is aimed at preventing progression. If behaviours can be modified prior to the onset of symptoms due to early interventions then the burden of OA is likely to be reduced.

Figure 1-1: X-ray data from patients with knee OA (grades 1 - 4) and acute inflammatory conditions of the knee joints [5]. *Notice in the images as the gap between the bones decreases as the KL grade is more severe. This image is from figure 2 in the article* [5].

The costs associated with OA are more than just financial. Many people who suffer with the disease state that its effects severely affect their quality of life. Many people often see that they are in increased pain for a large portion of the time and that, along with the functional impairment that comes from the effects of the disease, leads to a reported decrease in the quality of life of sufferers [6]. Depression is four times more common among people who experience persistent pain compared to those without [7] and around 20% of OA sufferers report experiencing depression [8]. Using the quality of life score (EQ-5D), a standard measure for health status which asks about their health status using five distinct measures relating to mobility, self-care, usual activities, pain/discomfort and anxiety or depression, scoring between 0 and 1, people suffering from a long term muskoskeletal condition, including OA, had an average quality of life score of 0.58 compared to those without a long-term condition who had a score of 0.92 [8]–[10].

Across several countries, it is clear to see that there is an increasing cost because of OA [11], [12]. Although there has been no real study into the UK expenditure on treating OA, several other countries have looked into the expense of treating the disease [13]. The cost of OA is not just measured as the cost of drug and treatment, direct costs, it also includes indirect and intangible costs [14]. Indirect costs include factors like days off work, loss of productivity and any benefit payments a person may receive. Intangible costs are things like the cost of pain and suffering and the decreased quality of life along with the added risk of potential to develop anxiety and depression. The intangible costs are a point of controversy as their value can be perceived as different for every person [12]. There is no clear analysis of the real costs of OA; however, any reported amounts are likely to be

significantly less than the actual costs. Across the UK, the 2005-2006 average for the overall cost of topical and oral nonsteroidal anti-inflammatory drugs (NSAIDs) was £8.5 million and £25 million, respectively. The projected 2010 price for the drugs after adjusting to inflation was £19.2 million for topical and £25.65 million for oral NSAIDs [15]. Surgery, both arthroscopy and joint replacement, also have a substantial cost to the National Health Service (NHS), with hip and knee replacements estimated to cost £850 million [14], [16]. The impact of OA on the UK economy is huge, with an estimate of the total cost being 1% of the gross domestic product (GDP) [17], and the social cost of OA in Spain has been reported between 0.25% and 0.5% of the country's GDP [18]. In 2002 the Department of Work and Pensions estimated that during that year 36 million working days were lost because of OA, resulting in £3.2 billion in losses relating to economic production [19], and OA is currently costing the UK economy £2.52 billion annually through 25 million lost working days [20]. Across the globe the costs of OA are among the highest in the healthcare arena. In the USA in 2013 OA was recorded as the second most expensive condition requiring treatments, with the hospital costs reaching $16520 million [21]. Even without a definitive measure to the economic cost of OA, it is clear that the costs are significant, and they will continue to rise. The need to find out how to prevent the disease onset and potentially reduce spending is increasing as there are younger people being diagnosed with this disease so without change to implement early interventions spending will surely increase.

The third largest area of NHS spending in 2013-2014 was musculoskeletal conditions, including OA costing £4.7 billion [22], [23]. By 2017 the total cost of osteoarthritis and rheumatoid arthritis in any joint on the NHS and the wider healthcare system was £10.2 billion [8]. As more and more people develop the disease, the costs attributed are going to increase and put further strain on the NHS. In the UK in 2017 there were 120,581 knee replacements and OA was the primary cause for 98% of these [8]. Therefore, there is a clear and definite need for diagnostic aids and models to indicate risk of disease onset and progression, as currently there are no predictive tools like the ones proposed and developed as part of this PhD.

1.1.2. Motivation and Clinical problem

OActive is an EU-funded research project, aiming to improve healthcare by transforming and accelerating the OA diagnosis and prediction based on more holistic features than just clinical measures. OA is not an easy disease to define, predict or treat so the OActive

project looks to make patient-specific OA predictions and interventions by using different models and big data analytics to better leverage the information in the data. OActive's mission is to find innovative ways to use data with the aim of better understanding the onset and progression of the disease and improving patient outcomes.

The work in this thesis does not use image analysis, but instead considers the applicability of predicting radiological KOA status from easier to measure factors and features from clinical questionnaires. The features include measures such as Age, BMI, and activity status. The work here would provide the base for a model that could be utilised as a screening tool. This would be useful as having a filter to help determine what candidates require further investigations, such as x-rays, would help to reduce the cost of diagnosis, and potentially help to speed up diagnosis, delay disease progression, improving the process from a patient perspective.

The main clinical problem is two-fold. Firstly, there is a need to determine who has KOA at their first presentation to a clinician. Then, of those without the disease, establish who is likely to progress to KOA after a period. By identifying those with the disease, it becomes possible to indicate which subjects require interventions, such as more frequent follow-ups, to assess how the disease is affecting them. Similarly, by highlighting individuals at risk of developing the disease it would be possible to offer actions that may allow for a reduction in risk of early onset. This may be help to lose weight, reducing the BMI of an individual, taking them from a high risk to a low-risk group for developing KOA in a five-year timeframe.

Although there is a clear clinical need for tools to help diagnose and predict the risk of KOA, the models developed and described in this thesis are a preliminary step toward a version that could be used within NHS clinical practice. The models developed use data from the US where the population demographic is different from the UK, which would be the target audience from these models. In order for these models to be used in the UK they would require validation on UK based data to ensure the features are still relevant given the different population demographic. From this point, the options relating to modelling for the UK would be to remodel entirely on the UK data or to use multitask learning to enrich the data sources with the aim of producing a more generalised model, suitable for both demographic cohorts.

1.1.3. Review of related research

1.1.2.1. Interpretable Methods

As machine learning (ML) systems become more embedded in applications throughout the real world, a need for regulations has become apparent. Legal regulations for the systems are being developed, with different levels of restrictions being placed on the amount of risk they pose when used. As many ML models consist of complex structures [24], [25], there is an urgent need for those models to be interpretable and explainable [26]. However, this urgency is not consistent across all domains, but is a certainty in medical applications, such as diagnostic decision support, where it is crucial to understand what the model is doing and how the predictions were calculated.

ML models are used in a wide variety of application domains, with methods typically put into place for experts in the application area to interpret and understand the results in a way that makes sense to other people. In the real world, for areas that machine learning methods are being used it is critical for the successful and appropriate implementation and safe crossover that mathematical algorithms that are used in decision-making processes are capable of being integrated into human reasoning models. Importantly, the provision of interpretation for AI is now arguably a central pillar for the "right to explanation" written into the General Data Protection Regulations (GDPR) which came into force on 25th May 2018 [27], [28].

Machine learning methods may be interpretable by design, typically in the form of rules in induction trees such as Chi-squared Automatic Interaction Detection (CHAID) and Classification and Regression Trees (CART) [29]. Alternatively, rules can be derived from neural network models such as Orthogonal Search Rule Extraction (OSRE) [30], which uses the statistical properties of regularisation of non-linear statistical methods to smooth decision boundaries and so reduce the effect that noise in the data has on the extracted rules. In contrast, more parameterized and sometimes less interpretable models require model calibration in order to be used effectively. However, the importance of calibration of probabilistic models [31] often favours the use of logistic regression (LogR) and its non-linear extension, feedforward neural networks and typically the Multi-Layer Perceptron (MLP). The partial response network (PRN) is a method that has the performance benefits of the MLP with the interpretability of a linear model, such as logistic regression.

In this thesis, interpretability refers to the user's ability to extract meaning from the results produced [32]. One definition of interpretability is broken down into three criteria [33]. The first criterion is to assess if the explanations are immediately understandable. The next criterion is whether the results are relevant and align with knowledge in the area. The third criterion is the predictive stability. For example, when there are similar cases stability would be evidenced through similar predictions for those cases.

1.1.2.2. Applications of Machine Learning to Osteoarthritis Research

There are five stages of OA according to the Kellgren-Lawrence (KL) scale [3] where Stage 0 is classed as no OA and Stage 4 is severe OA present in the joint. A clinician usually analyses and classifies images for diagnosis. More recently, studies have been conducted using artificial neural networks (ANN) to predict if a patient has OA based on blood samples [34]. Using computer aided diagnosis (CADx) would provide an objective tool to support clinicians. This particular approach has the diagnostic accuracy of a human clinician. By using both humans and machines there is the potential for more reliable diagnoses [4].

Another approach being tested is using plain radiographs and ANNs to determine if OA is present and how severe it is. This study aimed to make diagnosis and classification more stream-lined as there is some uncertainty in the stages and several clinicians may disagree over the classification of a patient in the KL scale [35].

Image analysis using MRI scans and x-rays to aid in diagnosis are also key resources that can be used in data mining and knowledge extraction leading to more informed decisions. A task within the OActive project was focused on considering the segmentation of MRIs and extracting geometrical features. This work looked at data from the OAI project and the Zuse Institute Berlin and used a semi-supervised approach to implement the segmentation through a multi-atlas learning process. The resultant image mask can then be split into bone specific masks where a transformation for distance is calculated and clustering is performed. The results of this process are feature descriptors for each part, all consisting of the mean, standard deviation and a plot of distance values [36].

Image analysis is widely used in the healthcare arena, with research related to KOA proving to be no different. A 2017 paper suggests that by using image extracted data along with questionnaire data it is possible to use machine learning to generate predictive models for the incidence of KOA [37]. Another example is a modified type of random

forest models being used to predict KOA, where feature detection from images were determined through the use of a 3-stage constrained local model [38]. Also, there is an increasing trend to consider multimodal approaches in diagnostics, using both convolutional neural networks to analyse images and model stacking to combine the neural network findings with more standard clinical measures [39].

When identifying the work that has been done so far in the area of KOA it is apparent that much of the research has gone into trying to classify plain radiographs into the different KL stages [35], and trying to determine if automated systems, such as neural networks, have a place in medical diagnosis [40], [41]. One of the papers even looks at trying to show that complex models can be made more interpretable for use in the medical domain [42]. The majority of the papers mentioned here look at the end result – a patient with symptomatic knee OA, and only then the attributes that could have become risk factors for trying to determine propensity to the disease, and possibly slow the progression. OA is a key research topic as the cost to the UK as a result of OA is growing year on year, and that trend can only be assumed to continue in the future as more and more people are living longer, and are at a higher risk of developing OA [14].

The data used in machine learning studies referring to OA have been collected in two different ways. The Framingham Study had a sample of 1805 patients, with only 79% able to be studied. The origins of this data were in the Framingham Heart Study. The OA data consisted of x-rays and medical history of the patients. The Framingham study was the 18th Biennial Examination of a long term study that started in 1948 and concluded in 2014 [43], [44]. When looking at interpretability for precision healthcare the MIMIC-II data was used [42]. The data consisted of counts of medication, diagnoses, and lab tests for the 32289 subjects across two groups. For trying to determine the severity of knee OA using the KL scale two main datasets have been used. The Osteoarthritis Initiative (OAI) and the Multicentre Osteoarthritis Study (MOST) data have both been used, with OAI being used more than once. The OAI data contains information on 4796 subjects in the form of x-rays and clinical factors [35], [45]. The OAI data has information on other factors of interest, such as demographic and self-reported activity information, but they are not used in these studies. The extra quantitative data, and that the data is publicly available make this a good candidate for use in any initial exploratory investigation. The MOST is another publicly available dataset that has over 3000 subjects. One example that looks at diagnosing Nephritis and Heart Disease using a variety of inputs and targets, presented as binary data, use data collected from the UCI repository. The study into

nephritis had 120 patients, and the study for heart disease had 267 patients [40]. The final example of data that is used from an existing study if the 5th Korea National Health and Nutrition Examination Survey (KNHANES V-1) [41]. This data is made up x-ray data on various parts of the knee that are likely to show signs of the disease. The KNHANES V-1 is made up of 2665 participants. The other type of data, data specifically collected for new research was used in the 300 patients used to see if there was a way to determine OA from rheumatoid arthritis and both of those from 'normal' people without any arthritic disease. The data was made up of 38 inflammatory proteins present in the blood serum [34].

A method that has been used most frequently, regardless of whether image data or quantitative data was used is neural networks (NN). Artificial Neural Networks (ANN) are commonly used in conjunction with medical data. ANNs have the ability to distinguish between patient groups in a study [34]. This type of NN can be used in conjunction with logistic regression (LR). LR serves to determine what predictors are of clinical interest, in a type of feature selection, while the ANN is used as decision support [41]. When trying to automatically recognize the areas of the joint that are of interest Convolutional neural networks (CNNs) are used. Another example of multiple algorithms being used at once is support vector machines (SVMs) and CNNs. The SVM looks to extract the relevant information whilst the CNN focuses on the extracted details to try and identify areas of disease [45]. In one particular study Deep Siamese CNNs were used [35]. These specific NNs contain two or more identical networks and are favourable as there are fewer parameters to train as they have the same value in both networks. They also are able to handle class imbalances better than traditional NNs. Another approach used in different analysis is the use of decision trees and random forests (RF) [14], [42]. The advantage of decision trees is that they are relatively intuitive and therefore easy to interpret and understand. Random Forests are slightly more difficult to understand intuitively as they are an ensemble method. However in the study with RF there was also an interpretability model used (LIME locally – interpretable model – agnostic explanation) [42]. Looking at these approaches using LR as a baseline for determining risk factors for OA would be a good step, followed by the use of different machine learning approaches, varying in complexity and interpretability, to establish if there is a trade-off between the models ability to predict disease presence and how easy they are to extract information for use in a clinical setting.

Consistently, NNs performed very well at classifying knee OA, Nephritis (99%) and heart disease (95%). The RF/LIME approach rendered an 80% balanced accuracy and as importantly provided results that were consistent with the medical understanding of the illness. The decision tree method was determined to still require some further analysis as the criteria were not perfect, and giving everything an equal weight in the final decision is flawed for a disease that has risk factors that vary in degrees of severity. One prevailing result that was determined is that it is imperative to not treat the KL scale as definitive boundaries, but more as a continuous scale as the disease progresses in some areas faster, and more aggressively than other areas. ANN has been found to be a cost effective screening tool when used on x-rays and in conjunction with an experienced radiographer if the need arises. For example, one method for screening for knee OA would be to x-ray everybody's knees and have the images undergo the ANN algorithm to determine if there are areas of concern. Should the model the flag anything up that requires further investigation, bring these to the attention of a radiographer and a clinician to double check the results and follow up with a management, or treatment plan with the patient.

1.1.2.3. Research on Early Onset and Treatment

Clinical papers identifying risk factors are frequently based on a literature review of previous studies. The majority of these works looked at determining the suitability of risk factors for determining OA [46]. Behaviours, habits and lifestyles are all understood to be involved in some way with the onset and progression of OA [47]. Another paper looked at what OA is and which risk factors predispose a person to the disease [48]. Identifying advances in imaging and biomarkers was another area that some literature focuses on [49]. Studies concerned with the quality of life and societal impacts caused by OA were also literature based [11]. The other type of study is clinical trials investigating risk factors. One study tried to develop a prognostic model for knee OA [50]. Studies using publicly available datasets can be used to determine if problems such as alignment are a contributing factor into OA of the knee [51].

For literature review based studies, such as systematic reviews, a large volume of papers were screened, but it was common for only a small sample of these to be included. The three remaining studies that use data to derive conclusions about onset and early treatment use either study data or clinical data. The clinical data used was gathered from an experiment on mice, to determine the use for cathepsin B for early diagnosis of OA [52]. When focusing more on the risk factors a person may have population studies were used, namely the Rotterdam study, Chingford study, MOST and OAI [50], [51].

For one of the literature review-based studies a scoring tool was devised and used to assess the quality of the studies being used to determine the risk factors of interest [46]. The other literature review-based samples also followed a score-type system to assess the usefulness of the features found from the highlighted papers. The practical approach studies used different approaches. The study using the Rotterdam and Chingford data developed a risk prediction model to determine risk factors for knee OA [50]. The study utilising MOST and OAI data used analysis of alignment measures of the knee and interpretation of x-rays to determine the effect of malalignment on knee OA [51].

The majority of the papers report a similar group of risk factors; however, similar factors appear with different definitions, so features may be similar but not exact. In this thesis, there has been a need for matching of factors when using different data sources. One example of factor matching is in one dataset the reference is made to knee stiffness, whereas in another dataset the reference is knee pain in the past 30 days. These two are not identical but are similar and therefore can be matched for use in the modelling process. The literature searches brought in many of the same factors, proving to be consistent across what is currently used by doctors. These factors include increased BMI, previous knee injury, Hand OA and older age [46]–[48], [50]. Each specific study found additional feature that may put a person at increased risk of developing OA, for example one suggested that poor mental health and having had a hysterectomy were two risk factors for developing OA of the knee [46]. One paper suggested that features such as ethnicity, hormonal status, genetics and the presence of certain biomarkers may all lead to an increased risk of OA [48]. The conference on Osteoarthritis from 2000 also suggested that OA may not be a single disease, but a group that ultimately have a similar final common pathway, however this is unclear [48]. One suggestion for an early treatment would be to administer vitamins to people who may be at increased risk to reduce the oxidative stresses which can be responsible for inflammation in the joints, leading to knee OA [47]. Among early interventions, such as supplements and physical therapy some research suggests that the best indicator of future changes in the knee leading to OA developments are early radiographic changes which are commonly missed [50]. This is backed up in part by the study suggesting that a malalignment in the knee can speed the progression of OA, so early intervention to correct this would be a suitable approach for delaying progression, and potentially onset of OA [51]. Early detection tools are required to successfully intervene early enough in the disease to promote change in the outcome. One such method is the near-infrared fluorescence (NIRF) probe along

with cathepsin B, however further testing would be required before this approach could be applied to human cases [52]. A key finding is that is key to have treatments available for pre-OA conditions as a way to delay and prevent the onset of OA [49].

It is necessary to develop a concise set of risk factors that can accurately be attributed to early onset of the disease so that preventative and early intervention treatment measures can be used to help delay the progression and potentially prevent the onset of the disease. Taking these early steps would help to preserve the quality of life for people that is usually lost once the disease takes hold [11]. By having this type of model in place the economic burden caused by osteoarthritis on the whole would also be reduced. In order for this to happen, pre-symptom interventions are required.

1.1.2.4. Models for Risk Stratification

Risk prediction models are key for educating the public about their risk relating to developing a certain illness or disease. Many models exist for diseases that can cause great pain and suffering to those afflicted, such as cardiovascular diseases and cancers. Even though osteoarthritis is not a life-threatening disease, it does cause life altering disability, pain and suffering to those with the disease. Risk prediction models, in a medical application, have the ability to use information relating to a disease to calculate a person's chance of developing a disease over a given time period. Developing such a model for osteoarthritis would be key to helping reduce the impact this disease has, and therefore the costs associated with it.

At present, the only risk prediction models that exist for osteoarthritis are research based and used to demonstrate that such models can be developed for this disease. Many of the risk prediction models that have been developed so far make use of the well-known risk factors that are linked with the disease. Many of these models are then also useful for looking at the effect of modifying the risk factors to estimate the risk reduction [53]. Some investigation have been able to make use of varying datasets to develop a prognostic model for knee OA [50]. A key factor in any risk prediction calculator is that it is easy to use and understand, while retaining accuracy in the model [54]. The power of other risk prediction models is that they can be easily used by both clinicians and patients. Having a calculator that can be easily used by the patient gives them the power to make small changes at a point far earlier than a doctors consultation, as people are increasingly likely to research their own symptoms and try to improve them before seeing a doctor, aiding

in early management of the disease [55]. A more recent risk prediction calculator has been developed to include radiograph and MRI data [56].

When developing a risk prediction calculator, it is key to have enough data, without the chance of fitting to the noise present in the data. For full model testing, there should be the training data, used to develop the model, and at least one validation dataset that was not used in the development of the model. There are several examples of this within the papers mentioned. The first example is the Nottingham model that was developed on 424 participants from a cohort gathered from a hospital in Nottingham [53]. The first validation data used on the developed model was from the OAI, followed by a second validation dataset from GOAL. The use of multiple validation datasets helps to clarify that the model that is being used is suitable for use on unseen data and does not fit the 'quirks' present in the training data. Another example of multiple data use is the study from 2014 [50]. Two versions of the Rotterdam study were used along with the Chingford dataset to build test and validate the model. Other data that was used for the risk prediction models were the Fifth Korea National Health and Nutrition Examination Surveys (KNHANES V-1) [55], different selections of patients present in a hospital [54] and varying groups of the OAI dataset [56].

Developing interpretable models for use in healthcare is vital. It is important that decisions are made with clarity and that the processes that go into making decisions about a person's health are easily explainable to the patient and understood by doctors. For a long time, logistic regression models have been the models of choice in medical statistics for these reasons. Every variable, or individual risk factor, have a weighting that can be used to explain that features input into the model making the model one of the easier types to use in clinical circumstances. Several of these risk prediction models make use of logistic regression [50], [53], [56], one uses the less interpretable, but arguably more accurate artificial neural network [55] and one carries out probability analysis on features and predictors that are known to be influential in OA modelling [54].

Risk prediction models, even when used solely for research, have the power to help with new insights, and show the type of models that are able to be developed and used for helping individuals reduce their risk to a given disease, or at a population level to help promote change [53]. Much of the prediction power is to do with the data used in the study, and on the external validation datasets, for example, GOAL performed better than the OAI cohort on the same model, which may give information as to the way the dataset

was collected or the information it contains. Many risk prediction models for OA have only made use of easily obtainable information, such as simple biomarkers or data from questionnaires along with demographic information. One model showed that these extra information points offer little insight into risk prediction over what simple demographics alone can provide [50]. The biggest predictor into progression has shown to be minor radiographic changes, where interventions are still able to slow progression of the disease. The use and availability of risk calculators can help to educate people at risk of developing the disease on ways that they can reduce their risk. Providing people with a calculator that provides insight into the effect that interventions focused on risk reduction can have on their susceptibility to a disease is a powerful tool in both education and successful management of the disease [54]. Developing a model that utilises a more complex technique in a risk calculator resulted in a performance improvement compared with the more simple logistic regression [55]. Adding unnecessary extra terms in a model results in a model that is harder to understand, however in some situations adding the extra term may reduce error and can therefore be of benefit, especially in a situation where medical interventions can be the result [56].

1.2. Research Novelty

Within this thesis, several areas of novelty build upon existing ideas. The novel aspects are listed briefly below:

- Produce a diagnostic model based on all subjects from the OAI dataset who have sufficient data as defined in the study, where the presence of KOA is defined as a baseline score of KL 2+ at first presentation.
    - ► At present, the models that exist for determining the presence of OA are not specific to the knee and only consider age, joint pain, and joint stiffness. This model includes additional features and relates exclusively to the knee.
- Produce a prognostic model on subjects who, at baseline, do not have KOA (KL 0/1). This follows a longitudinal study for 5 years to identify subjects who develop KOA at KL 2+.
    - ► Currently, the only prognostic models available for KOA are for determining time from diagnosis to intervention, such as a knee replacement. The model described in this thesis looks at an at-risk individual and calculates the risk of disease in the next 5 years.

- The thesis includes a cohort based external validation for the risk models for diagnosis and prognosis of KOA. In both cases, the models developed in the analysis using the OAI data have been externally validated using the MOST dataset.
- Determine the suitability of the OAI and MOST data for multi-task learning with the use of the piling approach.
    - ► Existing MTL approaches consider the use of images in conjunction with clinical features, whereas this model solely relies on clinical features and data from multiple sources.

1.3. Thesis Overview

The work in this thesis details the modelling, both diagnostic and prognostic, applied to knee osteoarthritis, with the aim of developing models that can be used in clinical practice. The thesis chapters develop from initial modelling to validated models with app interfaces for easier use if implemented. The thesis also includes two chapters detailing work that build on the initial model but consider these in different ways. One chapter, Chapter 6, has the outlook to identify the influence of gender on the presence and development of disease. The other, Chapter 7, looks at how the use of a multitask learning approach can alter the model performance.

Chapter 2 outlines the data that is used throughout the analysis described in the thesis. The work detailed in the thesis utilises three datasets: Osteoarthritis Initiative (OAI), Multicenter Osteoarthritis Study (MOST) and OActive. The OAI and MOST data are both longitudinal studies that allow for diagnostic and prognostic modelling to take place. The OAI data is the primary dataset used throughout the thesis with the MOST data being used later, in Chapter 5, to validate the models. The OActive data only has one instance of outcome recorded for each subject, and therefore can only be used for diagnostic modelling. Similar to the MOST data, the OActive data modelling and validation is detailed in Chapter 5.

The diagnostic modelling using the OAI data from Chapter 2 is detailed in Chapter 3. The diagnostic modelling uses variables considered to be relevant following literature reviews of similar analysis [53], [56], [57]. This analysis considers the alternative use of different modelling approaches to assess which method has the best predictive performance and the most applicability to a clinical setting. In clinical practice interpretability and performance are both important for understanding the reasoning

behind model decisions being crucial when they have the potential to influence a decision regarding patient care. Using a pool of variables, a prediction about if a subject has KOA at the baseline visit is calculated and performance assessed using numerical and statistical methods.

The next step following on from diagnostic modelling is prognostic modelling, which is detailed in Chapter 4. Once again, this approach is carried out using the OAI dataset. The variables used to consider the future development of the disease differ slightly from those to detect the disease at the point of medical intervention. The work in this chapter considers different follow-up windows to have the most useful clinical impact and highest patient satisfaction. The model developed in Chapter 4 can be used to educate patients about their risks to developing KOA and can help to provide tips as to how to change their behaviour to reduce their risk.

The next step for this analysis is described in Chapter 5. This incorporates the diagnostic and prognostic modelling, and the process of externally validating them with other data, namely the MOST and OActive datasets. The externally validated models are then developed into web based applications, such as those from the NHS for BMI calculation [58], Diabetes UK for finding out the risk of type 2 diabetes [59] and the QRISK3 risk calculator to calculate a person's risk of having a heart attack or stroke in the next ten years [60]. The apps would then be available for easy integration with current clinical practices. The apps could be used to help with clinical decision making about signposting of patients and for patient education about the condition.

The final two chapters take the original models and build on them with different aims. Chapter 6 has the aim of identifying how gender contributed to the risk modelling or knee osteoarthritis. We do this in two ways, first by considering each gender separately with its own pool of variables, initially with the original variables, and then with variables that include gender specific features. In the gender specific analysis variables such whether a female has undergone a hysterectomy are included to establish if there is a link between the presence or development of KOA and the gender specific features for the given sample we have. Finally, Chapter 7 looks at the utilisation of the multitask learning piling approach in a preliminary analysis to determine if the inclusion of more data when modelling would have any improvement when assessing the performance of the model.

Chapter 2: Data Preparation for Analysis

2.1. Introduction

Throughout this thesis, three different datasets are used for the different analysis conducted. For the diagnostic modelling the OAI, MOST and OActive datasets are all suitable to use as they have data collected at the subjects first presentation. The prognostic modelling is only suitable for use with the OAI and MOST datasets because they are longitudinal studies, as there are repeat visits in varying frequency, allowing for multiple follow-ups.

Different datasets used in analysis can render different performance, due to the data collection process and criteria for involvement in the study. The different datasets each contain data that has been collected from different centres. More granular detail relating to the OAI, MOST and OActive datasets are presented in the appendix on page 210. Having multicentre data to use when modelling, and validating a model helps to improve generalisation. This means that caveats due to a specific centres data collection technique are less likely to influence the performance of the model used. Different centres help to reduce the likelihood of a model overfitting to a single centres data. Here, where all of the data used is from different centres, the chance of overfitting is greatly reduced.

The decision was taken to use the OAI data for both the diagnostic and prognostic models. The other datasets are used to validate and assess the performance of the models. It is vital for models that are to be used in clinical settings to have undergone validation with an external dataset [61].

2.2. Osteoarthritis Initiative Data

The data primarily used in this analysis is from the Osteoarthritis Initiative (OAI) [62]. The data is available for public access at https://nda.nih.gov/oai/.

The Osteoarthritis Initiative is a multi-centre study, conducted over a 10-year period in America starting in 2004. The OAI dataset consists of 4796 patients at baseline, conducting follow-ups at 12-month intervals for 9 years, with follow-ups either in person at a clinical visit or via a telephone interview. The dataset is made up of subjects recruited based on their likelihood to develop knee OA. At the initial visit, there were a mixture of people who already had clinical KOA (KL2+) and those who had not yet been diagnosed. A list of inclusion criteria was advertised in various places for people to refer themselves to be part of the study, such as targeted mailings, local newspaper advertisements, and meetings in the community or local churches [63]. Figure 2-1 shows the advert that was used to recruit to the Osteoarthritis Initiative.

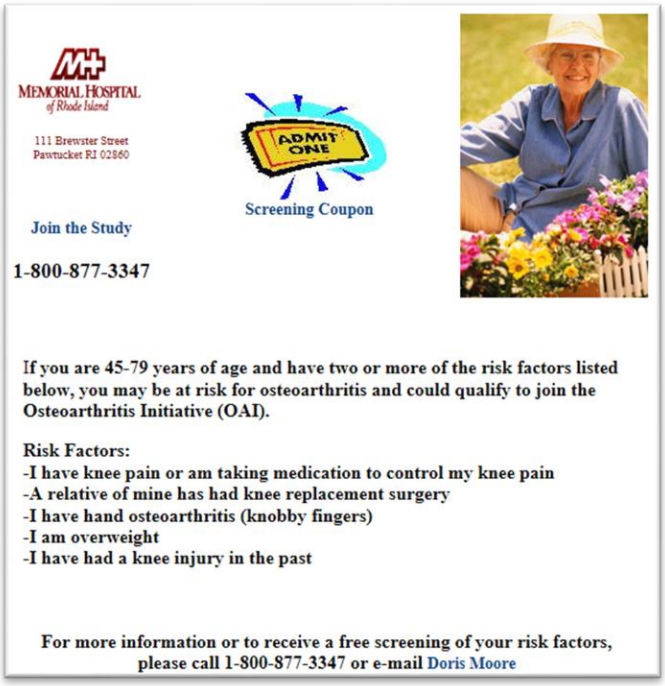


Figure 2-1: The advert used to recruit people into the OAI Observational study.

The OAI study protocol aimed to collect data on around 5000 participants, with roughly equal numbers of males and females. The ages of those included spans from 45-79. In the data collection, all ethnic groups were eligible for inclusion. The OAI study cohort is made up of three primary sub-cohorts: Progression, Incidence and Control. The progression cohort are defined as suffering from symptomatic OA at the initial assessment. The incidence cohort have the characteristics that would place the subject at increased risk of developing symptomatic OA during the study period. To be categorised as the incidence cohort, depending on age, other criteria need to be met, as shown in Figure 2-2. The control cohort will be made of a group that does not meet the requirements for the other cohorts. The exclusion criteria for the OAI study includes having rheumatoid arthritis, having had or plans in the next three years to have a total knee replacement, being unable to undergo MRI scans, a positive pregnancy test or any reason for not being able to provide blood samples [63].

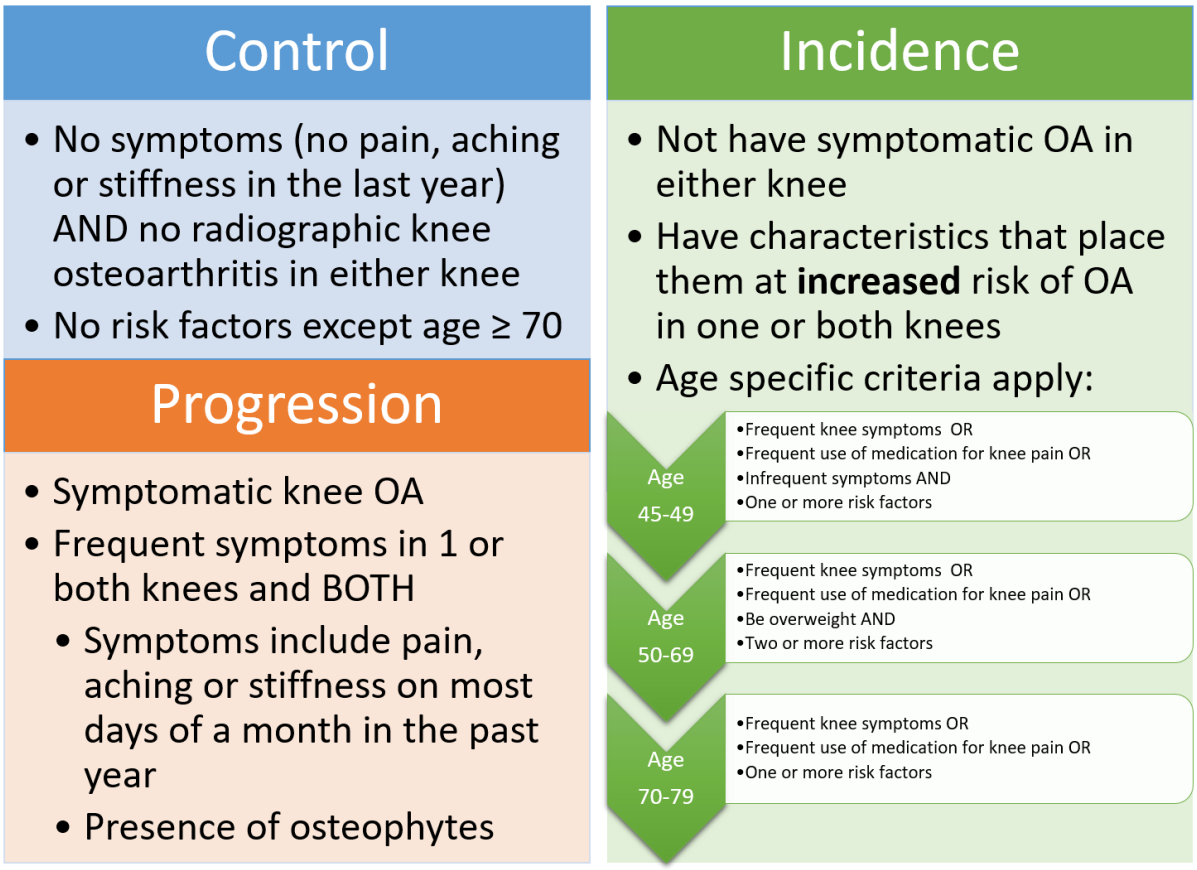


Figure 2-2: Criteria for inclusion in each sub-cohort, as defined in the OAI Protocol for the Cohort Study.

In the study, clinical examinations, questionnaires, and telephone interviews were conducted at varying intervals and the results were recorded. For the covariates used in the diagnostic analysis, only the primary recordings, taken at the baseline assessment, were required, but the data for follow-up visits were collected from other time intervals in the study period [64].

2.2.1. Data Pre-processing

In this thesis, in order to be included as a participant in either the diagnostic or the prognostic study, the subject is required to have a KL outcome recorded in the data; otherwise, they are removed from the analysis. The way in which this is defined differs for the two approaches and is explained in sections 2.2.2 and 2.2.3 respectively.

The process to gather the variables is consistent across both sets of analysis. The steps are shown in a diagram in Figure 2-3.

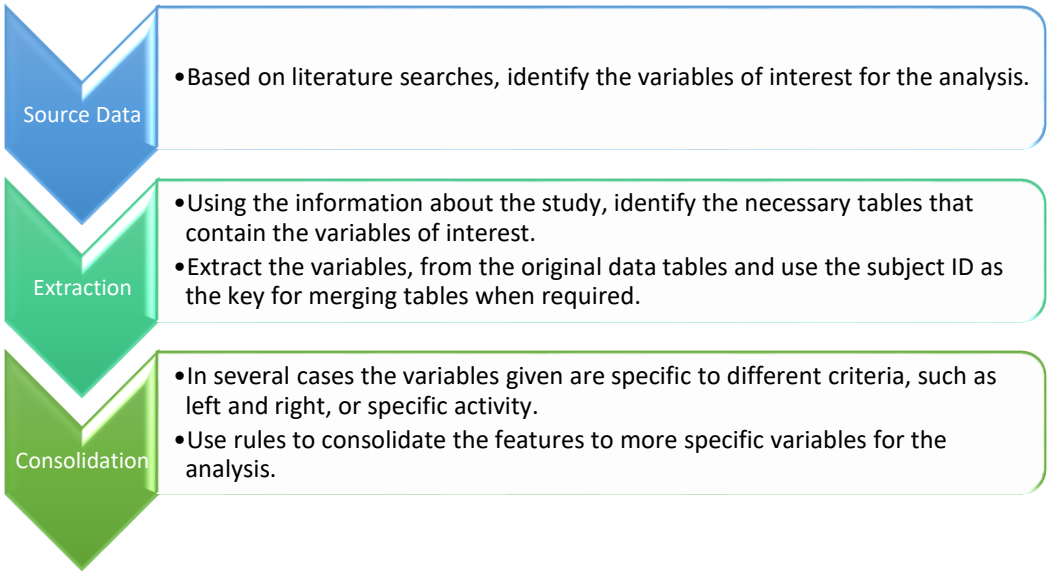


Figure 2-3: Stages to data pre-processing for the different datasets in the analysis.

The third step described in Figure 2-3 is Consolidation. This refers to where there is more than one variable in the raw data that fits the feature of interest. As the models discussed later in this thesis have been developed with the aim of being developed into clinical decision support tools, it was important to understand the way questions are worded, before any data manipulation could occur.

As many models used to give advice about symptoms rely on information from the subject's worst presentation of that issue, it was logical to arrange the data with the most severe option as the choice when more than one measure is available for a single subject. For example, where the questions ask if the subject has ever had an injury to their right knee, and then separately to their left knee, this information would be consolidated to a new question: have you ever had a knee injury, where the options are either yes or no. This would later ensure that the modelling matched the wording of questions presented to potential users, ensuring the most optimal subset of data or the modelling was used.

The OAI datasets provide detailed information on 4,796 subjects, with a large number of features considered under specific conditions for similar items and separately for left and right sides of the body. In cases where data is available for left and right hand sides of the body separately, then the most severe measure is selected and that is the measure for that specific feature, stored in a created variable defining the original left and right data.

The other type of variable creations is where there are several inputs to a created variable then the creation of that variable is considered in context with what the variable covers. This type is used where a variable may consider different activities individually, like sports

for example, or where family history is required, and this will make use of data from parents and siblings, consolidating it into a single measure. In this circumstance, with the question *'has family member x ever had knee surgery?'* the 'x' can be replaced by mother, father, sister, or brother. In each case the answer can be 'yes', 'no' or 'NA'. The rules to determine an answer to the question *'is there a family history of knee surgery?'* are that if any single question is 'yes' then the overall is 'yes'. If all answers are 'no', then the overall answer is 'no'. If any individual questions are 'NA' but others contain the yes' or 'no' options, then the 'NA' is ignored and the other rules apply. However, if all answers are 'NA', then the overall answer is 'NA'.

As the analysis conducted uses complete case, any 'NA' values in the final data, once the outcome variables are included, will be removed.

2.2.2. Diagnosis Cohort

For the diagnostic modelling, the original cohort had a sample size of 4796. After reducing the sample by removing those who have no KL grade, there remains a sample of 4507 subject. Finally, removing those subjects who have missing values in any portion of the variable sets leaves a usable cohort of 2707 subjects in the complete case analysis, as shown in Figure 2-4.

The data from each subset of variables, clinical and demographic, subjective and physical activity questionnaire variables, all underwent the same data pre-processing, and to merge the data for the pooled data, used in the analysis mentioned in Chapter 3, the subject ID was used. As each individual cohort has varying numbers of subject with KL scores, the loss of subjects between 3309 and 2707 is from missing values in the other variable subgroups, as not every subject has information relating to each variable of interest.

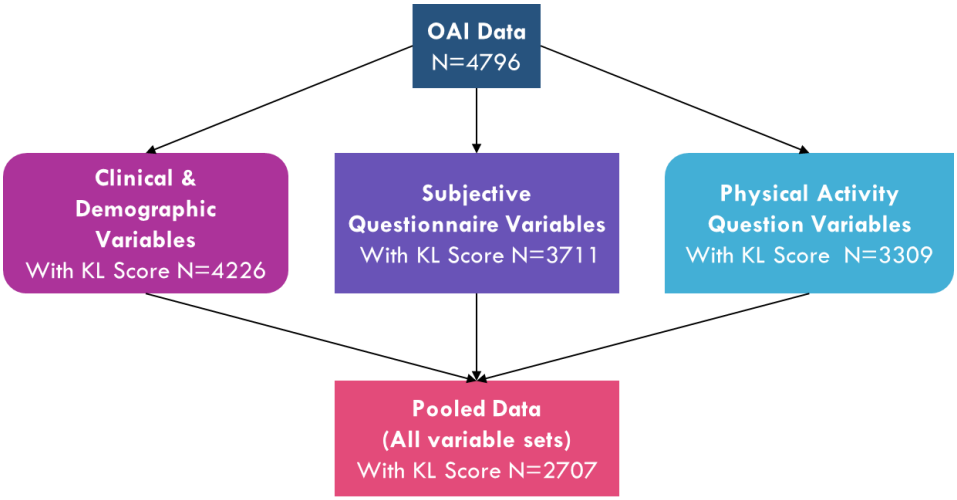


Figure 2-4: A visual representation of how the data required for the analysis was selected, and how it is made up of the data cohorts.

2.2.3. Survival Cohort

For the prognostic modelling only subjects with no baseline KOA could be included, as it looks to consider the time to change state from no disease to active KOA. Removing any subjects that had KOA at the baseline assessment left a sample of 2510 subjects. These subjects had no OA, in other words, a KL score of 0 or 1 at baseline.

When considering time to onset of disease the only subjects that can be considered are those with at least one follow-up measurement. Using this to filter, the sample size is reduced to 2314 subjects. In order to see the impact of given features on the likelihood of developing KOA, a set of covariates are also added to the outcome data. Considering basic demographic features for subjects where there are no missing values, the usable subject cohort is comprised of 2136 subjects. Figure 2-5 shows the way that the data has been pre-processed for the usable cohort suitable for the survival analysis.

The variables used in the prognostic model pool are the same as those initially considered for the diagnostic model cohort. The features selected for use in the model are explained in Chapter 4 in the results section.

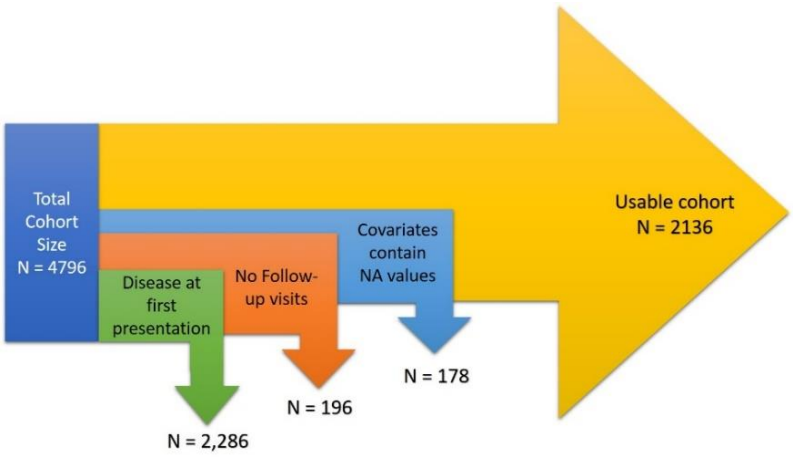


Figure 2-5: A Sankey diagram to visually display the data pre-processing.

2.3. Multicenter Osteoarthritis Study Data

The Multicenter Osteoarthritis Study (MOST) is a longitudinal, prospective, observational study of knee OA in older Americans with OA, or those who are at increased risk of developing it [65]. The data gathered in the study comes from two separate clinical centres, one looking at data coordination and one focusing on analysis. The MOST dataset enrolled 3,026 study participants and conducted five follow-ups at months 15, 30, 60, 72 and 84. At each follow-up x-rays were collected, except at the follow-ups for month 72 as these were telephone interviews only.

The community-based sample of 3026 men and women was made up of subjects aged 50-79 drawn from the general population. The selections were made in a way so that they are likely to either have pre-existing OA or be at high risk of developing OA as indicated by weight, knee symptoms or a history of knee injury or surgery to the knee, as shown in Figure 2-6.

To recruit to the MOST study, the two centres involved used a variety of methods. One approach used was targeted mass mailing. The centres compiled lists to send mail to who fell into the target age group. The lists were made from different sources including voter registration and health membership organisation (HMO) membership [66]. Another way that was used to recruit was community promotion, where the centres identified groups and agencies that had older clientele and targeted advertising to these. The final approach used was mass media. This is where the centres worked with local media in the areas to inform of the study and request volunteers.

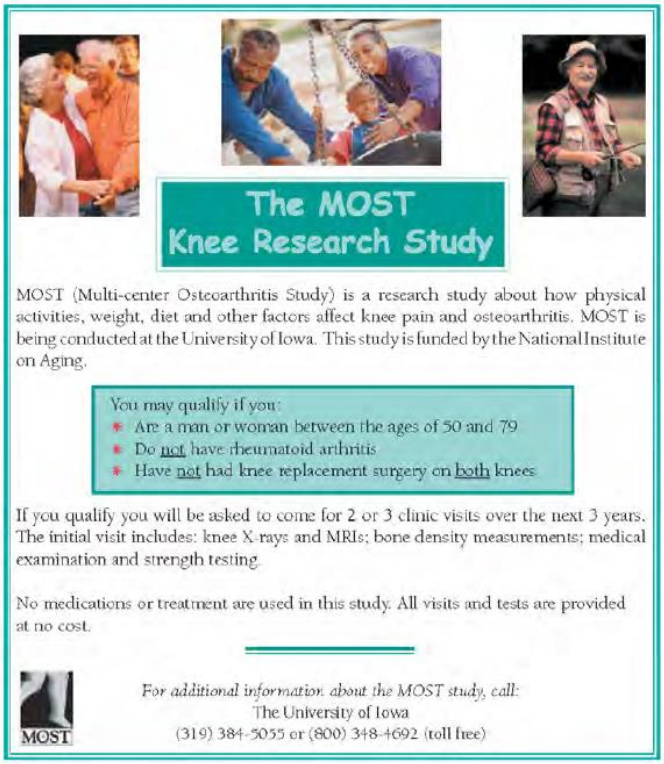


Figure 2-6: The advert used to recruit people into the MOST study.

2.3.1. Data Pre-processing

Depending on the analysis, the outcome measure will change and the way the cohorts are created will change. The pre-processing of the data is consistent with the approach illustrated in Figure 2-3. In the way that the OAI data was consolidated, the same approach was used with the MOST data.

2.3.2. Diagnosis Cohort

The diagnostic cohort requires subjects to have an initial KL outcome, and complete case information for the variables present in the model. Of the original 3,026 subjects 2,006 subjects have complete case information.

Of the variables used in the diagnostic model, the variable describing knee swelling is not present in the OActive dataset. To get around this, for the data to validate the OAI models, the predictions need to be marginalised.

To marginalise over the data, the first step is to filter the subjects through the combinations of binary variables in the model. Then, the OAI training data predictions are averaged for each variable combination that was filtered. The corresponding averaged prediction is then assigned to each subject that has matching covariates. Due to not all

combinations in the MOST data being represented in the OAI training data, after marginalisation the remaining MOST sample with predictions is 831 subjects.

2.3.3. Survival Cohort

Following the same process for the OAI data, all subjects with KOA at baseline were removed from the sample. Then ensuring that there was the baseline assessment measure and at least one additional follow up with a KL grade outcome resulted in a sample of 1190 subjects. It can be seen in Figure 2-7 that the number of events in the whole cohort is 404, where there is a large (786) number of censored cases.

The variables used in the prognostic model are different from those in the diagnostic model, but are consistent with the variables used from the OAI data to develop the model. This is further explained in Chapter 5. However, as the aim is a complete case study, then after removing missing values in the covariates there would be less than 100 subjects to test the model. To ensure that there is a sensible amount of data to validate the model, the OAI training data was used to calculate the mode value, which was then used to impute for the 'NA' values. The decision to use the OAI data for the imputation helps to prevent and limit any potential data leakage, specifically as the model was built with the OAI data. Imputation was only required for three variables; family history and history of falling imputation both assign 'no' to missing values and WOMAC score imputation fills 'NA' values with a score of 8. Imputation of these features happens only because the columns for those variables are missing. The imputation is only required as the MOST data is solely for the purpose of model validation, built using the OAI data.

Once the imputation for 'NA' values was carried out, the resultant seven-year and five-year cohorts contain 1,178 and 1,155 subjects respectively.

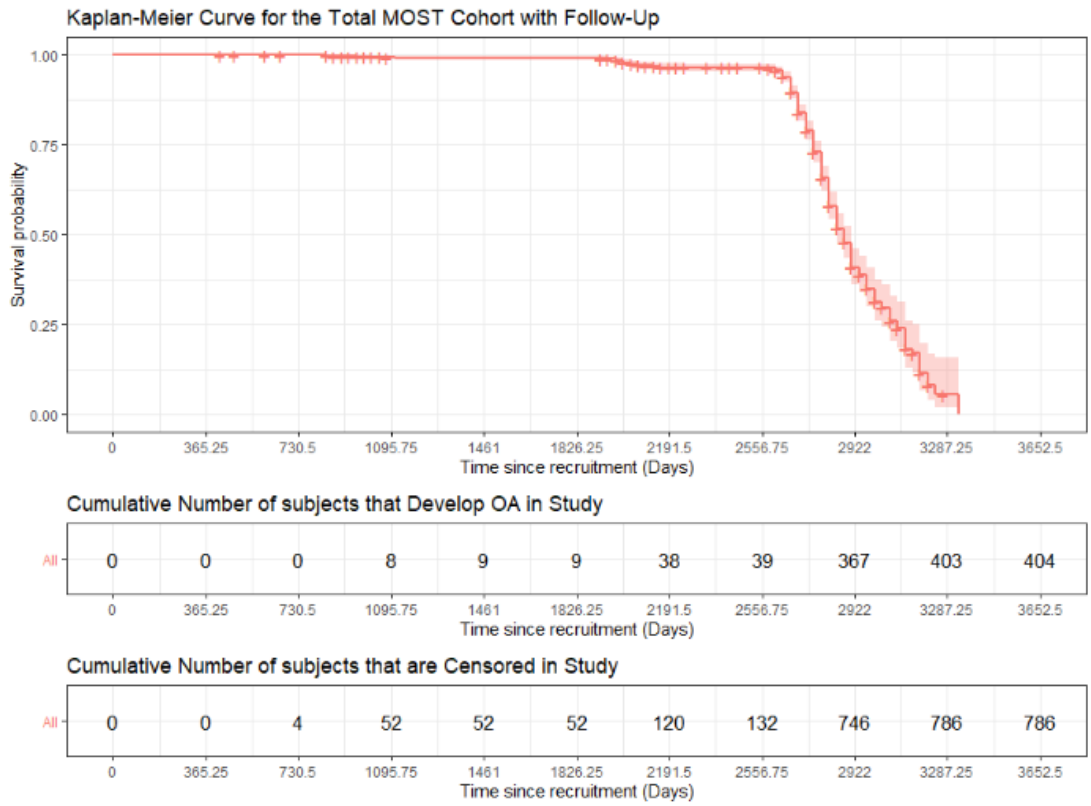


Figure 2-7: Kaplan-Meier curve for the whole population, with a table detailing, at yearly intervals, the cumulative number of events and censoring.

2.4. OActive Data

The final dataset used throughout this thesis is from the OActive project [1]. The OActive study is a multi-centre study, conducted in several centres in Europe. The data has been collected from three centres, in Greece, Cyprus and Spain. The centre in Spain focused on looking at healthy subjects who were at an increased risk of developing KOA, with the aim of trying to recruit over 100 subjects. The Greek centre focused solely on the evaluation of athletes who had suffered some level of trauma to the knee, aiming to recruit in excess of 90 subjects. This population was mainly younger than the age typically associated with the development of KOA. The final centre in Cyprus focused on elderly people with developing KOA, the population that is typically thought of when referring to KOA sufferers. This centre had the goal to recruit at least 130 subjects.

Figure 2-8 details the individual centres contribution to the way OActive is developed. The aim was to have a sample of at least 300 subjects recruited to OActive, however after removing missing values the usable dataset contained 206 subjects, split as shown in Figure 2-8. The majority of the OActive sample is comprised of elderly subjects, all of whom have KOA. The healthy at risk population are predominantly in the no KOA category (n = 71 from Spain, total n = 76), with only 5 subjects from that centre and

population type having KOA. The final centre, from Greece, only contributes 2 cases to this dataset, with 1 having the disease and the other not. The difference between the recruitment number and the usable sample number is due to the criteria for inclusion in the analysis, mainly age at least 45 and a given KL grade, and data completeness, with the removal of individuals with missing values.

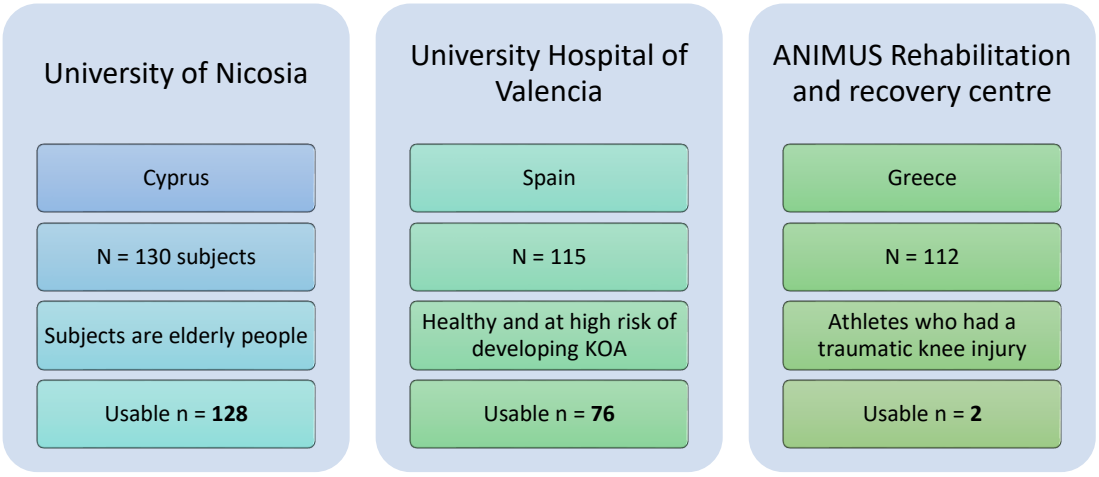


Figure 2-8: A graphic to shoe how the OActive dataset is made up. The OActive dataset is made up of 357 subjects recruited based on them meeting inclusion criteria defined by each centre. The centres in Greece, Spain and Cyprus recruited 112, 115 and 130 subjects respectively. The usable cohort sizes from Greece, Spain and Cyprus are 2, 76 and 128 subject respectively once the inclusion criteria is applied.

The data contains tables with information relating to biochemical and biomechanical measures, demographics, pain scales, social participation, and self-reported clinical measurements. The information contained in the dataset has the potential to give insights as to how osteoarthritis affects people, both in their everyday lives and biologically. The insight into how the disease effects people overall and not just clinically could better enable clinicians to treat people with a more holistic approach, instead of the typical pharmacological way that has been used as standard in OA treatment [67].

2.4.1. Data Pre-processing

The demographic information from the people in the study include healthy subjects who are at a high risk of developing osteoarthritis and elderly subjects who are likely to already have the disease. After following the steps illustrated in Figure 2-9, the usable cohort is 206 subjects.

The sole purpose of the OActive data in this thesis is to validate the diagnostic model within the parameters of the original OAI data. For this reason, only subjects aged 45 years and over are included within the validation data set. In addition, this helps to validate

any model developed using the OActive data as the other two datasets, OAI and MOST, exclusively look at those aged 45 and over, and 50 and over respectively.

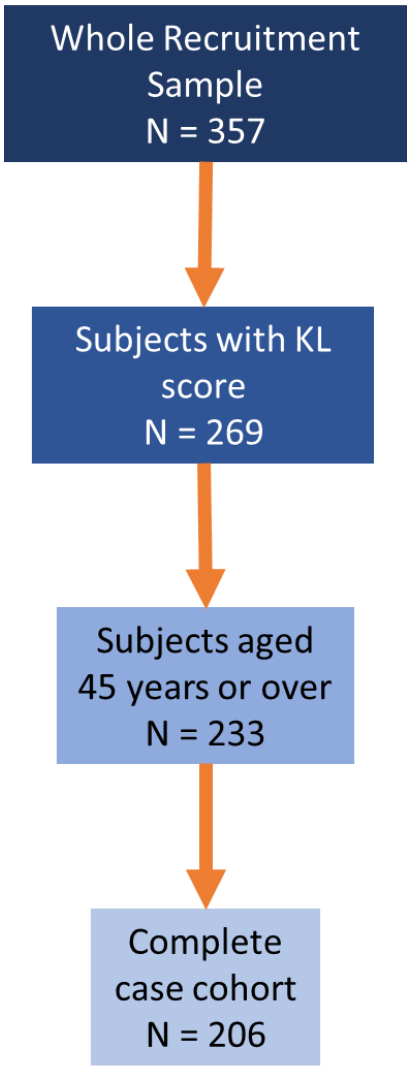


Figure 2-9: Diagram to show how the pre-processing of the OActive data is carried out.

2.4.2. Diagnosis Cohort

The diagnostic cohort requires subjects to have an initial KL outcome, and complete case information for the variables present in the model. Of the original 357 subjects, 206 subjects have complete case information.

Of the variables used in the diagnostic model developed with the OAI data, there are three variables not present in the OActive dataset. To get around this, for the data to validate the OAI models, the predictions need to be marginalised in the same way as for the MOST diagnostic cohort.

To marginalise over the data the first step is to filter the subjects through the 40 combinations of binary variables in the model. Then, the OAI training data predictions are averaged for each variable combination that was filtered. The corresponding averaged prediction is assigned to each subject that has matching covariates. After marginalisation, the remaining OActive sample with predictions has 206 subjects.

2.5. Limitations with the Data

Data from the real world often has missing values. There are many approaches to dealing with missing values, each having its own merits and disadvantages. There are two primary ways of dealing with missing values – deletion or imputation [68]. Commonly used approaches include imputation with forward and backward filling, multiple imputation, and complete case analysis. Another approach is to use analytical methods that can deal with missing values [69].

Missing data causes issues when modelling, such as making handling the data and analysis difficult, reducing efficiency in models and introducing bias [70]. In any method of dealing with missing values there is bias introduced into the data, so the type of imputation used in any given analysis may be chosen for what works best for the given dataset [71].

Imputation uses the available data to fill in missing values. Although this is a commonly used approach, forward and backward filling are known to increase bias and potentially lead to false conclusions as data will artificially have repeated measures, and are not often recommended. Mean substitution replaces missing values with the mean of that variable, without altering the sample mean for the variable. However, mean imputation reduces the correlations involving the variables that are imputed. This approach has some good points for univariate analysis but poses problems if considering this approach within multivariate analysis [72].

Multiple imputation, most commonly multiple imputation by chained equations (MICE), is designed for missing at random data but can be extended to cases where data are missing not at random [73]. However, MICE can encounter problems in data with a large amount of observations and complex features like nonlinearities and high dimensionality. It also poses the additional problem of being difficult to implement, where single imputation and complete case analysis are easier to implement [71]. In any imputation method there is the potential for data leakage, which can impact the way the data performs in models and can impact the accuracy of predictions.

The final approach for dealing with missing values is complete case analysis [74]. This is the most common way of dealing with missing values. Complete case analysis works by removing cases where there are missing values present; as a result, this approach reduces the sample size. One disadvantage of this approach is that if the data are not missing completely at random then removing instances with missing data will introduce bias [75].

Within this thesis, two main solutions to the problem of missing data are used: complete case analysis and, on a much smaller scale, mean imputation. By choosing to use a predominantly complete cases analysis, there is a systematic reduction in the type of bias added into the initial models, with the complete case data forming the basis of all modelling. As several of the modelling approaches employ this approach when missing data is present in the modelling set, there has been no real impact on the data that would be used, despite a large reduction in the size of the dataset.

The cases where mean imputation was used is solely for the validation work. This is due to an inconsistency with the available variables in the dataset to those already present in the model. This allows the use of two datasets that would not have otherwise been compatible for model validation with the OAI model. The validation is further discussed in Chapter 5.

2.6. Variable Selection

The diagnostic modelling uses variables considered to be relevant following literature reviews of similar analysis [56], [57], [76] . We know from the literature that features such as gender, genetic disposition, BMI and history of injury are all factors that contribute to the onset of KOA [77]. The variables used to consider the future development of the disease differ slightly from those to detect the disease at the point of medical intervention. Seventeen variables fitting clinical and demographic features were identified using the extracted OAI data. The variables include the age, gender, and BMI of the individual, along with information of family history, previous injuries, and diagnoses of osteoarthritis in other joints and general arthritis in the body. Several variables in the OAI data are self-reported. The self-reported data is made up from subject's answers to questionnaires relating to their symptoms and how they are impacted, recorded at the first presentation meeting, along with data made up of answers on questions about how much they take exercise and how this impacts them. An initial analysis looking at only the clinical and demographic work was conducted, and the subsequently presented at IDEAL 2019, showing the idea for the diagnostic model [78].

The justification for the inclusion of features in both the diagnostic and prognostic model revolve around a known risk to KOA. The risk of KOA increases as age increases, similar to BMI, both of which are present in the diagnostic model, and only BMI used in the prognostic model. Gender was another feature with a clear link, such that females are more at risk of KOA than males of the same profile, resulting in this variable being used in both models. Stiffness and swelling are known symptoms that can indicate the presence of KOA, resulting in these features being used in the diagnostic model. Mobility was considered in the diagnostic model in the form of difficulty getting upstairs and knee pain that limited activity in the prior 30 days. A reduction in mobility is an indication of increased risk of KOA. Family history of OA was included in the prognostic model as the potential to indicate a genetic link for future development of KOA. It was important to consider injury when considering future risk of KOA resulting in the variables for 'ever injured knee' and 'history of falling' as the former indicates a known risk whilst the latter may suggest a higher likelihood for injury, increasing the risk for developing KOA in the future. Finally, WOMAC was used in the prognostic model as it provides the self-perceived view of the condition from the subject's perspective, proving an indicator into how they feel at that time which may influence how that individual behaves in the future. For example, a high WOMAC score indicates a poor self-perceived view of the condition, possibly providing insight into how the person feels in areas of their life not covered by the other features.

2.7. Split-Sample Modelling

When modelling and assessing performance the goal is to determine the model performance in the general population. However, due to the nature of studies and data there is only a sample. To measure the performance of a model there needs to be separate training and test sets to assess if the performance is consistent for both. There are several approaches that allow for this comparison: bootstrap, leave-one-out and split-sample validation.

As with all options, there are positive and negative points that contribute to the choice of a method to implement.

Bootstrapping is a method in which a proportion of the data is held for testing while the data size is maintained though random duplication of samples. The training set is then

used to develop the model and the hold out set is used to test the model performance. This process is repeated with different test subsets a number of times. An illustration of this approach is shown in Figure 2-10. This produces a wide variety of performance values. This is useful as it can be used for calculating confidence intervals and standard errors. However, there are drawbacks to bootstrapping. One such negative is that the approach is not good for calculating the expected value as the predictions can often be inconsistent. The results may also depend heavily on the sample used to model. There also may be issues relating to sample size when splitting the samples when trying to ensure that the training sample is of a sufficient size prior to random duplications. Another problem is due to the duplicates present in the sample, leading the model to produce bias predictions [79]. Finally, bootstrapping is a computationally expensive method that obscures the explainability of the other approaches when applying context to the feature contributions to the outcomes produced by the model.

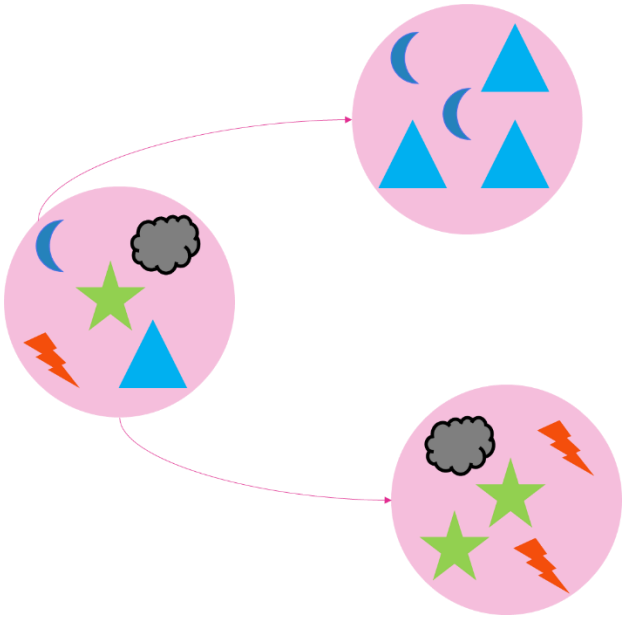


Figure 2-10: An example of bootstrapping. There are five observations in the first set with two subsequent sets created using the sample size equivalent to the original; however, in the other two sets there are repeated observations.

Leave-one-out is an example of an exhaustive cross-validation approach. This approach works by removing one sample for the test set and using the remaining samples to train the model. This is then repeated for each instance in the data, resulting in each data being the test set exactly once, shown in Figure 2-11. This approach, similar to bootstrapping, is useful when calculating the confidence interval and the standard error [79]. One of the key drawbacks is that, depending on sample size, the process may be computationally expensive and may still contain bias due to the modelling samples.

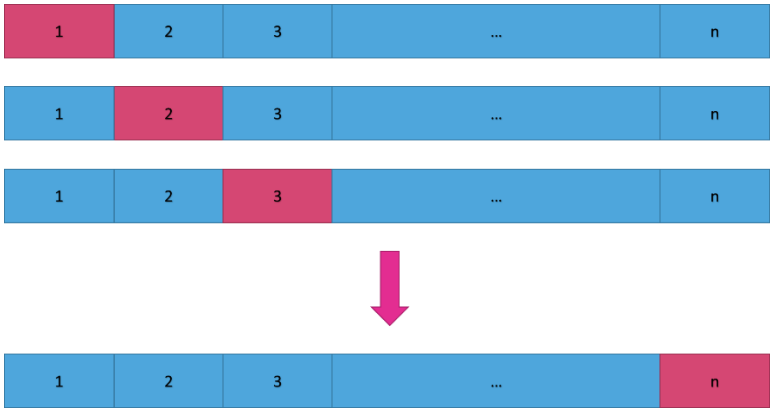


Figure 2-11: A visual representation of the leave-one-out approach to model validation. The block in pink shows the hold out set, whilst the blue samples are those used to train the model. This system allows each sample to be the test set exactly once.

The final approach of validation that was considered was split sample. In split sample validation the dataset is split into test and train subset, where the train set is used to develop the model and the test set is used to assess model performance. An example of the data split is shown in Figure 2-12. This approach has the advantage of using a single model, which is vital when building a model for explainability as this can be used to show and explain what factors are relevant to the outcome. This approach is also advantageous as the data size for modelling is large and for testing will likely produce a representative outcome due to the large sample [79]. As the data size is not small, a resampling approach is not necessary here. However, the key drawback is that by chance the test set split may not represent the general population, therefore producing bias results.

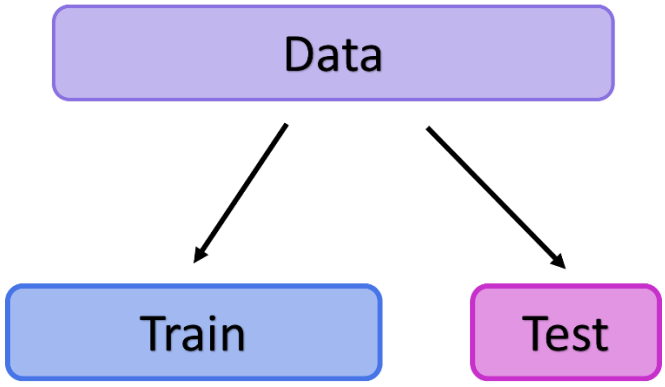


Figure 2-12: This is a visualisation of how the data is arranged when using the split sample approach for model validation.

The work in this thesis uses the split-sample approach to model validation. This is due to the way in which decisions can be explained, as one of the key themes throughout the work in this thesis in interpretability. This split sample approach also allows for the assessment of the generalisability of the model whilst also avoiding overfitting the data to the model.

2.8. Discussion

The models developed in Chapter 3 and chapter 4 are done with the OAI dataset. This was the publicly available data that could be acquired at the start of the project. The models built with the OAI data are validated using the OActive and MOST datasets. These are detailed in Chapter 5.

The data are primarily used in a complete case format, with some imputation by marginalisation used for the validation datasets. The analysis conducted in this thesis in Chapters 3, 4, 6 and 7 rely on complete case data to build the models, with the work in Chapter 5 making use of complete case for modelling and imputation by marginalisation over the available variables for the validation dataset. This approach was chosen, as the algorithms were the point of interest, interpreting the data given to provide results. Each imputation method adds bias, but the complete case analysis is easy to implement and straightforward, giving reason why it is the most popular method when dealing with missing values, despite its disadvantages.

Chapter 3: Diagnostic Model for Propensity of Presenting with clinical KOA at baseline assessment

The research defined in this chapter has been published in the Lecture Notes in Computer Science book series and was presented at IDEAL, Manchester, November 2019 conference.

McCabe PG, Olier I, Ortega-Martorell S, Jarman I, Baltzopoulos V, Lisboa P. Comparative Analysis for Computer-Based Decision Support: Case Study of Knee Osteoarthritis. Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics). 2019.

Available from: https://link.springer.com/chapter/10.1007/978-3-030-33617-2_13

3.1. Introduction

This chapter describes the first step in the process of identifying disease in 'at-risk' subjects. This analysis focuses on predicting the presence of KOA at a baseline visit to a clinician without prior knowledge of whether the subject has KOA.

In clinical settings where patient education and behaviour modification are at the forefront, interpretable models are key, as it is vital to be able to explain a decision that leads to any course of action related to an individual subject. This chapter will look at several approaches to interpretable models applied to the Osteoarthritis Initiative data, and compare the performance of the models.

The aim of the work is to produce a tool that can identify presence of the disease from risk factors given an outcome, in this case the radiographic KL score, without the need for an initial x-ray. By using baseline values, the goal is to identify if it is possible to diagnose on first presentation without the need for anything more than questionnaire results, reducing the cost on the clinical practice and the anxiety of exploratory medical interventions for the subject. The aim is that this tool will be used as an aid to assist clinicians in the initial decision-making process for signposting patients to the correct services to best streamline their treatment in the most efficient way. This tool, however, would require further validation before it could be used within NHS clinical practice. The data for modelling is from the US, which fundamentally has a different demographic to that of the UK. Verifying the model is suitable for use in the UK is vital for the successful implementation of a model to clinical practice.

The use of machine learning (ML) models by application domain experts requires a method to interpret the operation and inference made by these complex methods, in a language that people can understand. It is critical for successful translation and to ensure safety of real-world applications, that mathematical algorithms are capable of being integrated into human reasoning models.

Clinical models need to be explainable, interpretable and easy to understand [80]. It is a crucial part of developing a model for clinical practice to ensure that there is a high level of understanding and interpretability within the model [81]. As with many areas, there is often an example that is the exception to the rule, but in healthcare there is the need for decisions about treatment to be transparent and explainable. In safety-critical applications such as clinical medicine, the combination of machine explanation and generalisation test comprise the steps of verification and validation that are central in software development

methodology. This framework is integral to regulatory frameworks which apply equally to the use of AI in decision making [31]. Crucially, the provision of interpretation for AI is now arguably a central pillar for the "right to explanation" written into the General Data Protection Regulations (GDPR) [82].

Predictive modelling currently exists for different diseases. Risk prediction models are key for educating the public about their risk relating to developing a certain illness or disease. Many models exist for diseases that can cause great pain and suffering to those afflicted, such as cardiovascular diseases and cancers [83], [84]. Predictive models are even used to tackle the problem of operating room delays, administering of antibiotics to new-borns and for treatment plans following hip or knee replacement surgery [85]. Even though osteoarthritis is not a life threatening disease, it does cause life altering disability, pain and suffering to those with the disease [13]. Risk prediction models, in a medical application, have the ability to use information relating to a disease to calculate a person's chance of developing a disease over a given time period. Developing such a model for osteoarthritis would be key to helping reduce the impact this disease has, and therefore the costs associated with it [14].

At present, the only risk prediction models that exist for osteoarthritis are research based and used to demonstrate that such models can be developed for this disease. Many of the risk prediction models that have been developed so far make use of the well-known risk factors that are linked with the disease. A lot of these models are then also useful for looking at the effect of modifying the risk factors to estimate the risk reduction [76]. Some investigation have been able to make use of varying datasets to develop a prognostic model for knee OA [50]. A key factor in any risk prediction calculator is that it is easy to use and understand, while retaining accuracy in the model [54]. The power of other risk prediction models is that they can be easily used by both clinicians and patients. Having a calculator that can be easily used by the patient gives them the power to make small changes at a point far earlier than a doctors consultation, as people are increasingly likely to research their own symptoms and try to improve them before seeing a doctor, aiding in early management of the disease [55]. A more recent risk prediction calculator has been developed to include radiograph and MRI data [56].

Developing interpretable models for use in healthcare is vital. It is important that decisions are made with clarity and that the processes that go into making decisions about a person's health are easily explainable to the patient and understood by doctors. For a

long time, logistic regression models have been the models of choice in medical statistics for these reasons. Every variable, or individual risk factor, has a weighting that can be used to explain that features input into the model making the model one of the easier types to use in clinical circumstances. Several of these risk prediction models make use of logistic regression [50], [53], [56], one uses the less interpretable, but arguably more accurate, artificial neural network [55] and one carries out probability analysis on features and predictors that are known to be influential in OA modelling [54].

Machine learning methods may be interpretable by design, typically in the form of rules in induction trees such as Chi-squared Automatic Interaction Detection (CHAID) and Classification and Regression Trees (CART) [29]. Alternatively, rules can be derived from neural network models such as Orthogonal Search Rule Extraction (OSRE) [30], which uses the statistical properties of regularisation of non-linear statistical methods to smooth decision boundaries and so reduce the effect that noise in the data has on the extracted rules. In contrast, more parameterized and sometimes less interpretable models require model calibration in order to be used effectively. However, the importance of calibration of probabilistic models [31] often favours the use of logistic regression (LogR) and its non-linear extension, feedforward neural networks and typically the Multi-Layer Perceptron (MLP). The partial response network (PRN) is a method that has the performance benefits of the MLP with the interpretability of a linear model, such as logistic regression.

When identifying the work that has been done so far in the area of KOA it is apparent that a lot of research has gone into trying to classify plain radiographs into the different KL stages [35], and trying to determine if automated systems, such as neural networks, have a place in medical diagnosis [40], [41]. One of the papers even looks at trying to show that complex models can be made more interpretable for use in the medical domain [42]. The majority of the papers mentioned here look at the end result – a patient with symptomatic knee OA, and only then the attributes that could have become risk factors for trying to determine propensity to the disease, and possibly slow the progression. OA is a key research topic as the cost to the UK as a result of OA is growing year on year, and that trend can only be assumed to continue in the future as more and more people are living longer, and are at a higher risk of developing OA [14].

The aim of the study in this chapter is to ultimately produce a tool that can indicate the likelihood of the presence of the disease from risk factors given an outcome, in this case

the radiographic KL score, without the need for an initial x-ray at the point of first presentation to a GP. The clinical relevance of this tool is to inform the screening process. This is explored in detail in Chapter 5.

There are three research questions with particular clinical relevance: i) initial diagnosis of OA at first presentation; ii) risk of developing OA among the cohort initially diagnosed as disease free; iii) characterizing progression through different stages among the population diagnosed with OA. This analysis is focused on the first of these.

The focus of this chapter is to model clinical OA at first clinical presentation, which could lead to the development of a model that could be part of a screening process. The model identifies modifiable factors that influence disease onset, which enables feedback to be provided to subjects at risk in order to change their behaviour to help prevent or delay onset of OA.

This analysis in this chapter compares and contrasts the different approaches with the primary focus on accuracy of discrimination and explanation of model inferences. It also looks at taking features from a subject to develop an interpretable diagnostic model. This model could then potentially have clinical uses.

Chapter aims

- Determine the most interpretable modelling technique for predicting the presence of KOA at baseline.
- Develop a model that could be used to classify KOA without an x-ray, for use as a screening measure.

3.2. Specifics of the data and variable cohort development

The data used in this analysis is from the Osteoarthritis Initiative (OAI) [62]. The full explanation of the data is in Chapter 2. A visual representation of how the variables were grouped for the analysis is shown in Figure 2-3.

Seventeen variables fitting clinical and demographic features were identified using the extracted OAI data following a literature search. The variables include the age, gender, and BMI of the individual, along with information of family history, previous injuries, and diagnoses of osteoarthritis in other joints and general arthritis in the body. This is

further discussed in section 2.2.2. The features gathered after consulting the literature form the initial pool of variables used in the analysis. An initial analysis looking at only the clinical and demographic model was conducted, and the work presented at IDEAL 2019 [78].

Following the modelling on the individual data cohorts the inclusion of more than one variable subset is beneficial. This exploratory analysis suggested that the use of a wider variety of information gave a better determination for the presence of clinical KOA in a subject when compared with individual variable sets that only focus on specific criteria.

Several variables in the OAI data are self-reported. The self-reported data is made up from subject's answers to questionnaires relating to their symptoms and how they are impacted, recorded at the first presentation meeting. Similar to the clinical data, many of the variables needed to be compressed in order to be suitable for analysis.

In a similar approach to the Self-Reported dataset, the Self-Reported Physical Activity data is made up of answers on questions about how much they take exercise and how this impacts them. This set of data on its own appears to be the most modifiable in terms of lifestyle changes that a person can make.

In this analysis, the pooled data formulation has been created using all three variable sets. This data consists of clinical, self-reported and self-reported physical activity measures that need to be consolidated.

The chosen approach is to combine the datasets with the Subject ID and utilise a feature selector on the, now larger, dataset. The selected variables would then be used in the machine learning models. Then, after running the models with the selected variables, measures to assess the performance would be calculated and compared with the individual variable sets to see if the inclusion of different variables contributed to a better performing model.

This approach takes all of the variables in the data to be included in a pool and, through feature selection, determines which variables will add the most information to the model at a given step. The stepwise inclusion of the three variable sets enables the model performance to be monitored in terms of how the addition of new data affects the global model when compared to the baseline of the clinical and demographic data only.

Figure 3-1 shows how the OAI data, considering only subjects with a KL outcome, are split into variable cohorts of interest and the final variables included in the dataset applied to the modelling approaches detailed in section 3.3.

To determine which variables would be selected from each of the subsets, a voting technique was used. The process of voting used the stepwise feature selection along with CART and CHAID to determine which features were to be included in the variable subset. If a given variable was selected at least twice over the different approaches, it could be included in the variable subset cohort. In CART and CHAID a variable is chosen when it features in the final tree when that given tree was built using all of the possible variables available within the data.

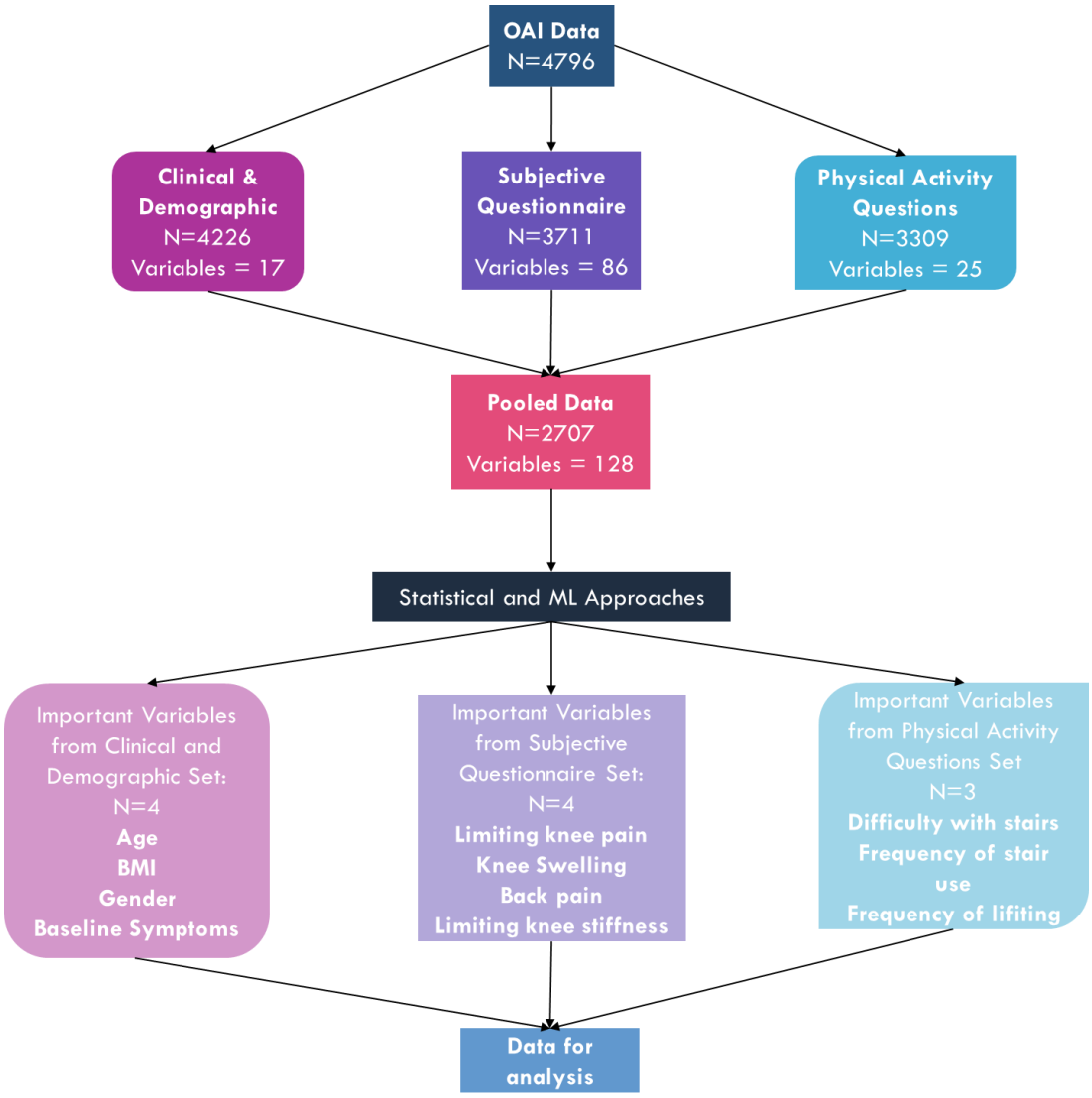


Figure 3-1: A visual representation of how the data required for the analysis was selected, and how it is made up of the data cohorts.

Of the 11 variables present in the final dataset supplied to the models, less are ultimately selected in each model through a process of internal feature selection. However, there is a cross-over in the variables determined as important according to each model when identifying the presence of KOA at baseline, based on these initial features.

3.3. Methods used

The methods used in the analysis performed here are machine-learning methods. Machine learning is an approach that can provide the ability to automatically learn and progress without being programmed explicitly. The type of machine learning used in these analyses are supervised machine learning. This is where there are previously labelled data that can train a model to make predictions given a set of predefined variables.

Throughout this analysis, four main approaches are used, and are detailed in the sections 3.3.1 - 3.3.5, with the performance metrics used in this analysis discussed in section 3.3.6.

3.3.1. Classification and Regression Trees

In this thesis, the Classification and Regression Trees (CART) were calculated using the R package, rpart [86].

CART is a rule induction approach that determines univariate cut points [87]. This machine learning approach can be classification or regression-based. In this decision-making, the classification approach is the most suitable option for the data, as the categories are clinical OA and non-clinical OA. In clinical situations, this can be used to develop a set of questions that can aid clinicians in a decision making process before invasive investigative tests are undertaken.

When trying to choose a machine learning approach to use a number of things are taken into consideration at each step. Many of the decisions are made regarding how easy the models are to use and understand. CART analysis has the advantage of being very interpretable and easy to understand. This is, in part, due to being able to represent the results in different form, such as graphically or with the tree diagram itself. Conditions to class membership are clearly explained meaning that the explanation about how the decisions are made are easily demonstrated, removing the 'black box' nature of machine learning. Another reason this approach is a favourite is due to the way the decisions made closely mirror those made by humans.

For this work, the focus was on classification of KOA status. Trees used for classification use binary splits, calculated using the Gini index, Equation 3-1 [88]. The Gini index

calculates the likelihood of a specific feature that is classified incorrectly when selected randomly. The value of this index is between 0 and 1, with 0 meaning there is purity of classification, with all elements deemed to be identical, or a value of 1, where there is a random distribution of elements across various classes, where the number of classes approach infinity. A Gini index of 0.5 means that there is an equal distribution of elements of some classes present.

When designing the tree, features possessing the least value of Gini would be preferred by the model. These features with lower Gini Index values are used for constructing the decision tree. The Gini index works in this instance as an in-built feature selector in the model. The index is a measure of total variance across the K classes. The Gini index is defined by:

$$Gini\ Index = \sum_{m=1}^{M} \sum_{k=1}^{K} \hat{p}_{mk}(1 - \hat{p}_{mk})$$

Equation 3-1: Formula for the Gini Index used then building CART models for classification.

Where $\hat{p}_{mk}$ denotes the proportion of training observations in the mth region that are from the kth class.

The value of the Gini index is small if the values of $\hat{p}_{mk}$ are all close to either zero or one. For this reason, the Gini index is referred to as a measure of node purity where a small value indicates that a node contains predominantly observations from a single class. The algorithm for growing a decision tree is an example of recursive partitioning. Each node is grown using the same rule set as its parent node and the data in the parent node is partitioned into its child node. The recursive process stops when conditions for bin size and complexity are reached.

3.3.2. Logistic Regression

In machine learning, logistic regression (LogR) is a type of parametric classification model. This means that LogR models are models that have a certain fixed number of parameters that depend on the number of input features, which output a prediction, which can be categorised by selecting a threshold value as a cut-off for binary classification.

In LogR the data is fit to an 's' shaped curve, called a Sigmoid function, Equation 3-3. This function takes a minimum value of 0 and a maximum of 1, which helps when the

goal is to classify samples in two distinct categories. By calculating the sigmoid function of the data, a probability of an observation belonging to one of the categories is produced.

Let $\Pr(y = 1|X) = p(X),\ X\epsilon\mathbb{R}, p(X)\epsilon[0,1]$,

$$\text{logit: } \log\left(\frac{p(x)}{1-p(x)}\right) = \beta x$$

Equation 3-2: Logit function, as the log odds ratio, of the chance an event occurs over the event not occurring.

Where β is the expected change in log odds of having the outcome per unit change in x, and

x is the value of the independent variable.

$$\text{Sigmoid function: } p(x) = \frac{1}{1+e^{-\beta x}}$$

Equation 3-3: The sigmoid function used in logistic regression.

This is the most commonly used statistical model in medical decision support [89]. Although it is linear-in-the-parameters, careful discretisation of continuous variables creates a piecewise linear model with the capability to model highly non-linear data, which are typical in clinical medicine. As a result, logistic regression models are often very competitive in discrimination accuracy compared with neural networks and other machine learning methods, except when interactions between variables have a significant role in decision-making, in which case rule induction may be preferred.

Logistic regression is a preferred method as it can also be translated into nomograms for easy clinical use and interpretation [90]. The use of a nomogram can turn otherwise complex mathematical models into easy to understand graphics that can show the real implications of changing behaviours to those seeking advice. For example, nomograms could be of particular use in educating a subject seeking medical advice how best to change their lifestyle in order to prevent developing OA or slow down their risk of progression. Equation 3-4 illustrates an expression where for binary covariates $\{x_i\}$ the exponentials show explicitly the size of the effect of the variable on the odds-ratio.

$$\frac{\mathrm{P(class|X)}}{1-\mathrm{P(class|X)}} = \prod_{i=0}^{n} e^{\beta_i x_i} = e^{\beta_1 x_1}.\, e^{\beta_2 x_2} \ldots .\, e^{\beta_n x_n}.\, e^{\beta_0 x_0}$$

Equation 3-4: Logistic regression odds ratio formula

Where n is the number of independent variables.

3.3.3. Lasso

Lasso (least absolute shrinkage and selection operator) is a shrinkage method used in statistics and machine learning to perform both variable selection and regularisation to aid in prediction accuracy and model interpretability. The shrinkage relates to the ability to discard variables that are not as useful in the model. This approach is preferred over subset selection as they are more continuous and therefore have lower variability. When used in conjunction with the partial response network (PRN), the lasso is used for variable selection [91].

Lasso uses L_1 penalisation, as in Equation 3-5. This means that by adding a penalty equal to the absolute sum of the coefficients the method will shrink some parameters to zero, so some variables will not play any role in the model. Using Lasso in this way is one approach to select features in a model. The penalty performs a continuous variable selection process in the model.

$$\hat{\beta}_{lasso} = \underset{\beta}{argmin} \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 + \lambda \sum_{j=1}^{p} \left| \beta_j \right| \right\}$$

Equation 3-5: L1 penalisation term

Where n is the number of observations, and

p is the number of variables in the data.

The λ term is the hyper parameter hat adjusts the penalty term. When $\lambda = 0$, no parameters are eliminated, and as λ increases more coefficients are set to zero and are eliminated. Lasso forces less important features to have a β value equal to zero, removing the feature from the analysis.

3.3.4. Multilayer Perceptron Automatic Relevance Determination (MLP-ARD)

A multilayer perceptron (MLP) is a type of artificial feed-forward neural network [92]. The MLP is made up of at least three layers: an input layer, a hidden layer, and an output layer. The output layer in this case is a binary classifier.

For the MLP-ARD configuration, a standard MLP is used in the first instance [93]. The automatic relevance determination (ARD) is useful when it is important to know what variables are contributing the most to the classification [94]. The ARD is to determine

the most relevant features in the data. The theory behind this is Occam's razor, which is a principle that states a preference for simple theories [95].

In machine learning a model that can leverage the same amount of information but containing fewer terms than a competing model will be preferred as there is a preference for a simpler model. If a model is too complex, it will fit well to the training data, fitting also to the noise and as a result will perform poorly on unseen test data.

$$\frac{P(\mathcal{H}_1|D)}{P(\mathcal{H}_2|D)} = \frac{P(\mathcal{H}_1)}{P(\mathcal{H}_2)}\frac{P(D|\mathcal{H}_1)}{P(D|\mathcal{H}_2)}$$

Equation 3-6: The ratio between theory 1 and theory 2. The ratio is how much the initial belief favour theory 1 over theory 2 and how well the given data were predicted by each theory when compared.

One common problem with neural networks is that they can tend to overfit to the data they are trained on. This can be partly rectified by using Bayesian approaches, as it can use Occam's razor to automatically infer how flexible a model should be given the data [96]. Therefore using this, for a network trained on data $D = \{X^{(m)}, t^{(t)}\}$ by adjusting the weights, w, to minimise the error function,

$$E_D(w) = \frac{1}{2}\sum_m \sum_i \left(t_i^{(m)} - y_i(X^{(m)}; W)\right)^2$$

Equation 3-7: Error function between the true and predicted values calculated in the network.

Where t_i is the true value, y_i is the predicted outcome based on the inputs, $X^{(m)}$, and the weights in the network, W.

The term in Equation 3-7 is based on repeated evaluation of the gradient E_D using backpropagation. When using weight decay the objective function is modified to give Equation 3-8.

$$M(w) = \beta E_D + \alpha E_W$$

Equation 3-8: Modified objective function, accounting for the errors caused by the data and the weights, in a linear form.

Where E_D is the error caused by the data, and $E_W = \frac{1}{2}\sum_i W_i^2$, favours small values of W, and decreases the tendency of the model to overfit to noise in the training data.

For the ARD approach to work, a separate hyper parameter α_i is assigned to each group of weights spanning from the i^{th} input variable. The hyperparameter is re-estimated through each iteration of the tuning process. At the end of the training stage any

hyperparameters with large values indicate that an input has little impact on the final model meaning that their weights will decay to values near to zero [97]. This highlights what features can be dropped from the final model.

3.3.5. Partial Response Network

The partial response network, PRN, is a method to open the black box approach of the MLP [91]. The end product results in non-linear univariate and bivariate partial responses from the MLP. When the performance of the PRN is compared with a fully connected MLP, there is usually performance improvements because of the PRN implementation. The bivariate responses come from modelling pairwise interactions in the network. Interactions are modelled up to pairwise, and all others are categorised under the residual modelled in the network. The PRN implementation mimics models of deep learning but offers the advantage of being highly interpretable, in a similar way to a LogR model. The functional form of the PRN is given in Equation 3-9.

Equation 3-9: Functional form of the PRN given by the statistical decomposition of the multivariate effects into components with fewer variables.

$$\begin{aligned}\text{logit}\big(P(C|x)\big) \equiv \varphi(0) + \sum\nolimits_i \varphi_i(x_i) \\ + \sum\nolimits_{i\neq j} \varphi_{ij}(x_i, x_j) \\ + \sum\nolimits_{i_1\neq\cdots\neq i_d} \varphi_{i_1,\ldots,i_d}(x_{i_1}, \ldots, x_{i_d})\end{aligned}$$

Where C is the class member ship label, C is the target, x is the input, and the partial responses $\varphi_k(\cdot)$ are evaluated with all variables held fixed at zero except for one or two indexed as:

1. $\varphi(0) = \text{logit}\big(P(C|0)\big)$
2. $\varphi_i(x_i) = \text{logit}\left(P\big(C\big|(0,\ldots,x_i,\ldots,0)\big)\right) - \varphi(0)$
3. $\varphi_{ij}(x_i, x_j) = \text{logit}\left(P\left(C\middle|(0,\ldots,x_i,\ldots,x_j,\ldots,0)\right)\right) - \varphi_i(x_i) - \varphi_j(x_j) - \varphi(0)$

The way in which the PRN works can be explained in six steps, shown in Figure 3-2.

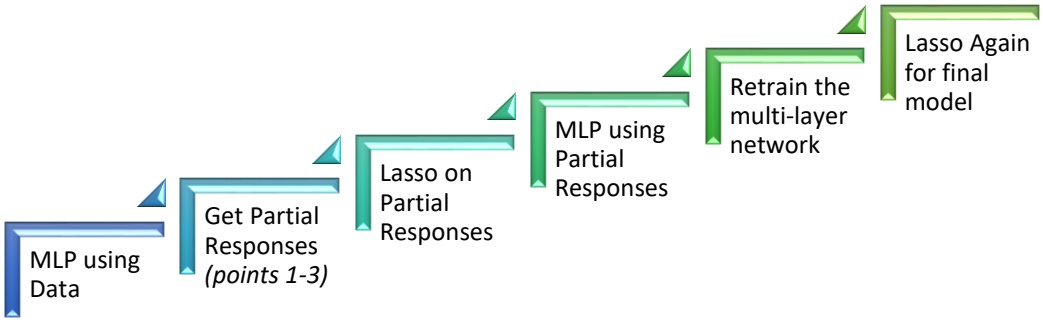


Figure 3-2: The six steps used to develop the partial response network.

3.3.6. Performance Metrics

For binary classification models a probability of belonging to a given class is calculated. In order to determine which class an individual belongs to, a threshold is needed. The threshold is usually determined by the prevalence in the population. As the population in these datasets who suffer from KOA is between 45 and 60%, the classification threshold in these models are set to 0.5. Therefore, a prediction less than or equal to 0.5 will result in a decision of no KOA, a negative instance, and a prediction greater than 0.5 would result in a decision of KOA, a positive instance.

The receiver operating characteristic curve (ROC curve) is a plot that graphically indicates the ability of a model to correctly classify binary outcomes as a threshold is altered. The area under the curve (AUC) is equal to the probability that a classifier will rank a random positive instance higher than a randomly chosen negative one [98]. In the AUC a value of 0.5 indicates a guess, with greater than this being deemed better than a guess, and lower than 0.5 being worse than a guess.

Sensitivity (Equation 3-10), specificity (Equation 3-11) and positive predictive value (PPV) (Equation 3-12) are all statistical measures of the performance of binary classification tests. The sensitivity measures the proportion of actual positives that are correctly identified. The specificity measures the proportion of actual negatives correctly identified.

At present, in the UK diagnosis of KOA typically follows an examination by a GP in conjunction with a consultation about symptoms to determine if the individual does have KOA, which is the gold-standard. There are no definitive tests for OA and x-rays are not always necessary, but tend to be used more when other conditions need to be ruled out, or for staging the progression of the disease, following an initial diagnosis. From the data

we have, as the subjects were part of studies related to KOA, there is no GP diagnosis, only a diagnosis that has been determined from examining the x-rays to establish what KL grade the subject has in their knee. As there are no definitive tests there can be cases where a diagnosis is missed, someone with the disease is told they do not have it, or someone without KOA advised they are a sufferer.

Sensitivity is the true positive rate; it is an indicator of how likely a model is to correctly identify a patient with a disease. If a model has high sensitivity it can help to rule out a disease where a person is not indicated to have the disease. Specificity is another measure assessing the way a model performs, this time indicating the true negative rate. This is a way determine how effectively a model can correctly identify people without a disease. Models with high specificity can be used to rule in disease in a person who is indicated to have said disease, potentially prompting further investigation. Positive predictive value, (PPV), is the odds of having the disease if you have a positive result. This measure is useful to both the patient and clinician as it can be used in conjunction with sensitivity to indicate how likely a positive result is actually true.

Equation 3-10: Sensitivity formula

$$\text{Sensitivity} = \frac{TP}{TP + FN}$$

Equation 3-11: Specificity formula

$$\text{Specificity} = \frac{TN}{TN + FP}$$

Equation 3-12: Positive predictive value formula

$$\text{Positive Predictive Value} = \frac{TP}{TP + FP}$$

Where:

- TP is true positive result,
- TN is true negative result,
- FP is a false positive result, and
- FN is a false negative result.

3.4. Results from Analysis

The variability in the model performance is, in part, due to the variables present in each of the models. As the models differ in complexity and variables present the model

performance will differ. One hypothesis is that there is a maximum possible predictive accuracy for any data set. This means that the preferred model is the simplest model that can perform at a comparable level to any other model, however complex, given the AUC. The results are summarised in Table 3-1.

CART has a PPV of 0.641 meaning that of the people who test positive for KOA only 64% will actually have the disease. The CART model will only correctly identify 60% of KOA sufferers. LogR has a higher sensitivity; this model would correctly identify 67.4% cases of KOA, and of those testing positive 61.3% will actually have the disease. The MLP-ARD has the lowest PPV, with only 56.7% of all positive cases actually having KOA. The PRN-Lasso has a PPV of 0.599 meaning that 59.9% of positive cases relate to a true positive result.

CART correctly identifies 77.6% of people without disease. LogR is next best at identifying people without disease, correctly identifying 71.6% of non-disease. The MLP-ARD is the worst at identifying non-disease with a specificity of 67.6%. The PRN-Lasso will correctly identify non-disease in 69.7% of cases.

Table 3-1: A table of performance metrics for the different models used in the analysis, giving the area under the curve (AUC), sensitivity, specificity, and positive predictive value (PPV).

	AUC	Sensitivity	Specificity	PPV
CART	0.719	0.600	0.776	0.641
LogR	0.763	0.674	0.716	0.613
MLP-ARD	0.778	0.677	0.676	0.576
PRN-Lasso	0.793	0.698	0.697	0.599

3.4.1. CART Results

The tree diagram in Figure 3-3 show the splitting criteria in a highly interpretable way that could be transformed into a question set for clinicians to use as a signposting tool. In the diagram, the middle number is the prevalence of people in that group that have knee osteoarthritis by following the conditions to arrive at that node. The bottom number is the percentage of the population that is covered by the node criteria.

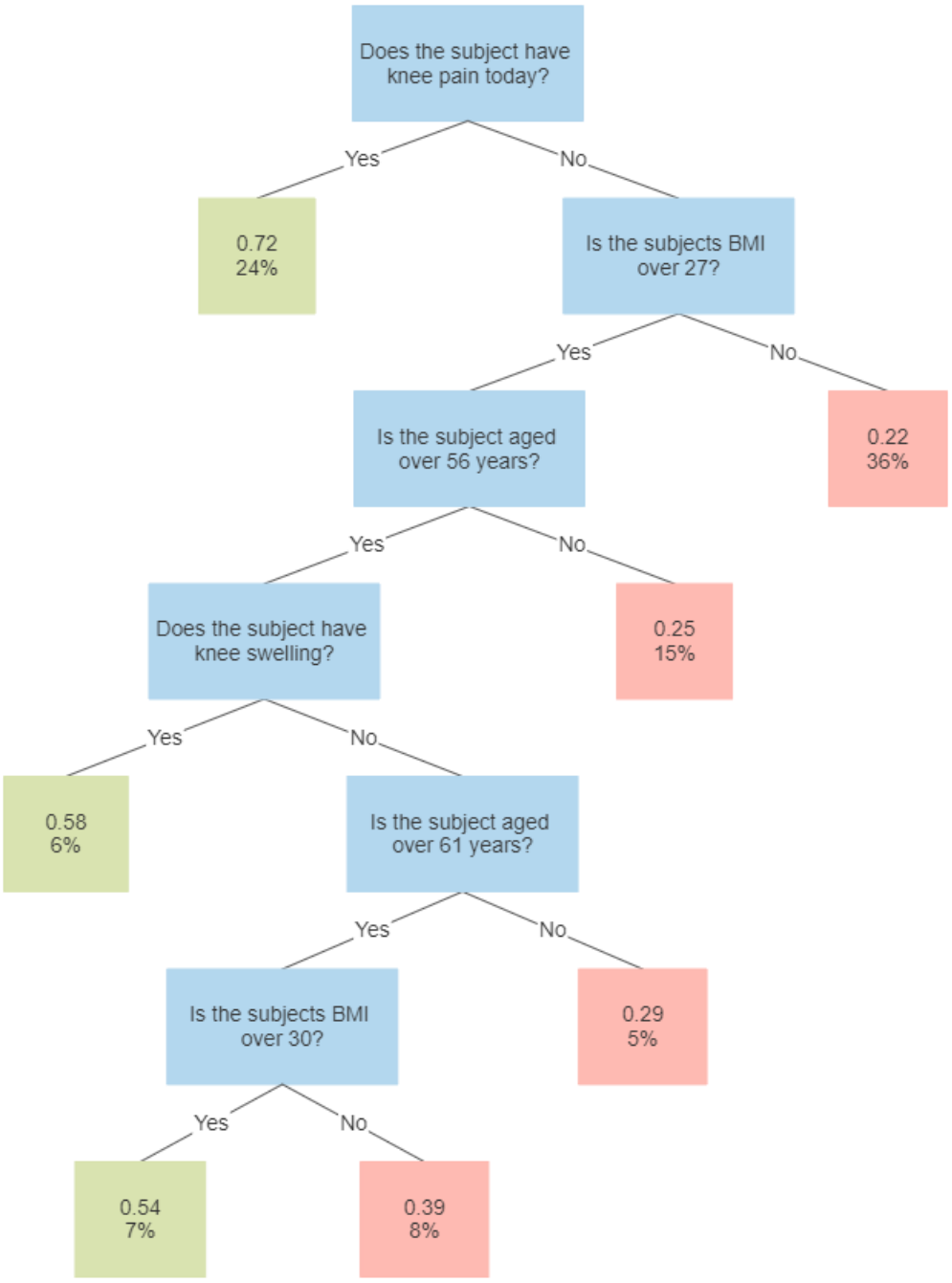


Figure 3-3: Tree diagram showing the stages in the process to determining the likelihood of the presence of KOA based on a set of questions.

3.4.2. LogR Results

The nomogram depicted in Figure 3-4 related to the performance of the LogR model. The LogR model is a baseline indicator as it is used in clinical practise, as this is the preferred method for binary classification for a variety of healthcare problems. The LogR model produces an interpretable nomogram that gives every value a point score that

relates to the probability of having the disease in question, in this instance, the disease is KOA. The nomogram also indicates a possible confidence interval where the symptom scores could fall, giving another reason that this type of approach is preferred in the medical arena. The results from the LogR analysis indicate that the data is well suited to this type of modelling approach.

The point score in the nomogram is used to assign a value between 0 and 100 to each predictor. To calculate the points the first step is to rank the predictors in order of the biggest to smallest impact on the model. The variable with the highest effect is then converted to 100 points, with the variables minimum state assigned 0 points. The remaining variables are then assigned a points value proportional to the size of the effect on the outcome [99]. This allows each state to be given a point score that maps to the odds of having KOA given a certain symptom set.

A nomogram ranks the importance of an effect in predicting the outcome only within the context of the other covariates in the model. It is important to also remember that the points do not reflect the association with the outcome but visually display the weight of the features considered when making a decision.

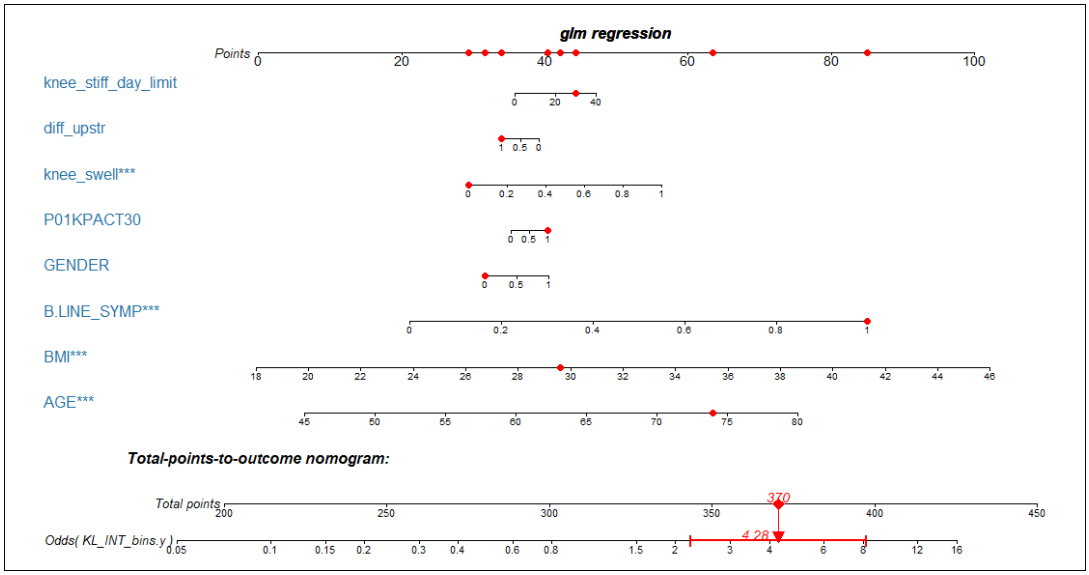


Figure 3-4: A nomogram produced from the LogR model indicating the odds of having knee osteoarthritis based on a set of features at different values.

3.4.3. MLP-ARD and PRN-Lasso Results

The calibration plots, shown in Figure 3-5, show how the model fit is near the pattern present in the data. This shows that at each step of the model, the model is adequately trained to perform predictions within the applicability domain. The models are well

calibrated as the points are all close to the diagonal line. The PRN-Lasso is the best-calibrated model.

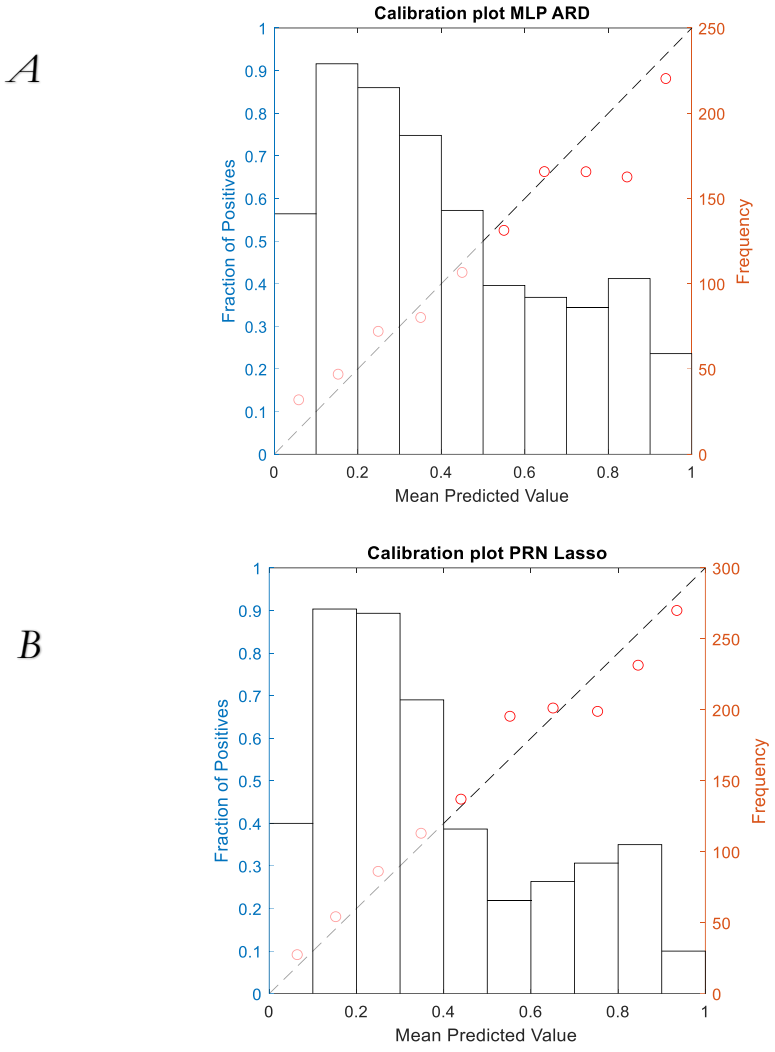


Figure 3-5: Calibration plots for the MLP-ARD (A) and the PRN-Lasso (B), showing how well the models are calibrated to the data.

3.1.2.1. Features after PRN-Lasso

After the initial MLP-ARD, the lasso model selects the most important features in the data. For this dataset, the main features are five univariate and six bivariate effects. The features still important after the PRN-lasso are the ones that are in the final model. In the final model, there are four univariate effects. These are: age, BMI, baseline symptoms (presenting with pain) and knee swelling. The effects can be shown in Figure 3-6.

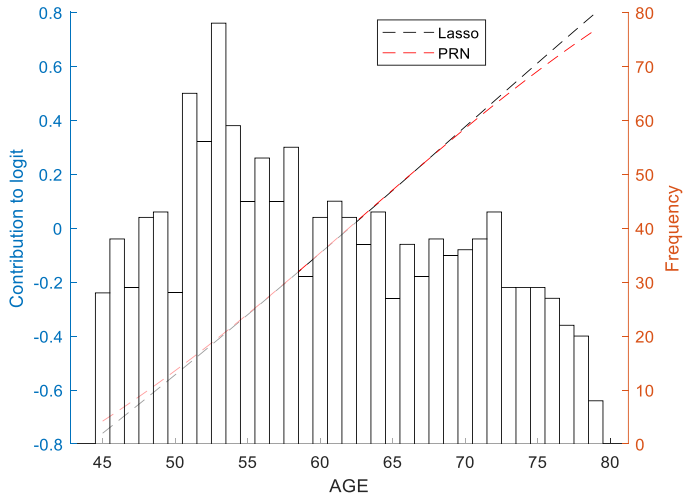
As age increases, the more the effect age has on the logit, so the contribution to the outcome, the presence of KOA, increases. The contribution to the logit is nearly linear until about the age of 70 where a similar pattern can be shown with the BMI and its effects to the presence of KOA. As the BMI increases, the contribution to the logit also

increases in a nearly linear pattern. The presence of pain symptoms at the investigation will increase the contribution to the logit. The subject experiencing knee swelling will increase the contribution to the logit as this symptom would indicate the presence of knee osteoarthritis. Both of these statements are in line with what is presented in the literature.

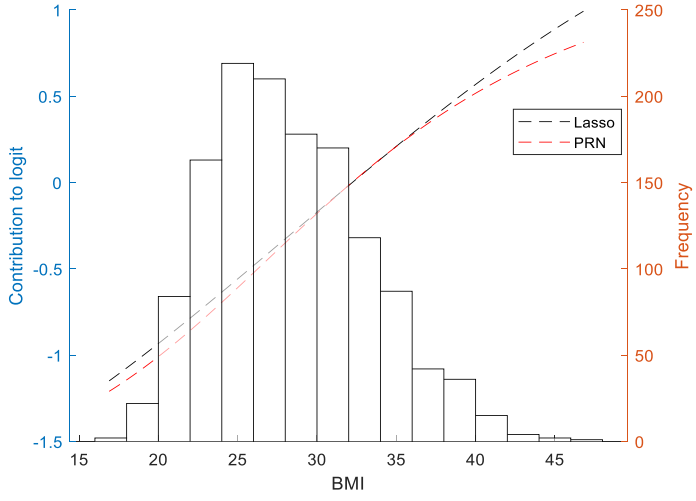
The AUC measure shows how well the models predicted the binary categories. These are shown in Table The models vary in complexity, with the most complex model being the MLP-ARD, and the most simplistic model is in fact the PRN-Lasso, as it only uses four of the eleven input variables, which are age, BMI, baseline symptoms (presenting with pain) and knee swelling.

The dependence of covariates explicitly is estimated using partial responses that are not constrained in any way. The result is that the functions are almost linear across the full range, as shown in Figure 3-6. This explains why LogR works so well. Namely, the assumption of linear dependence on covariates is met.

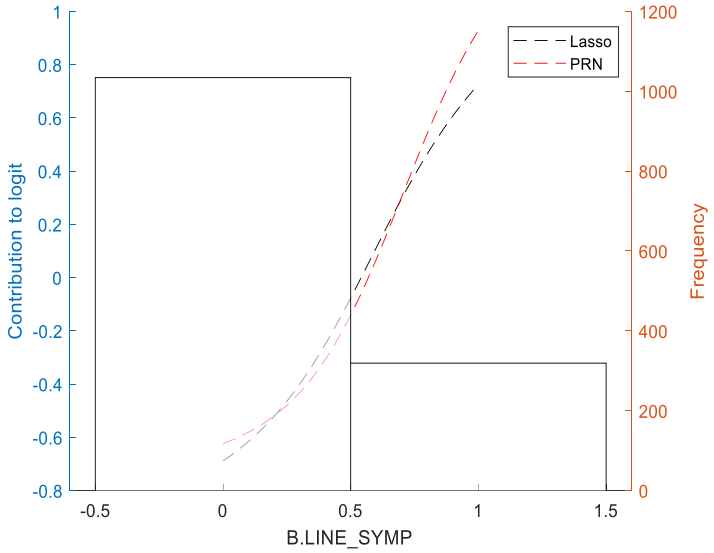
A



B



C



D

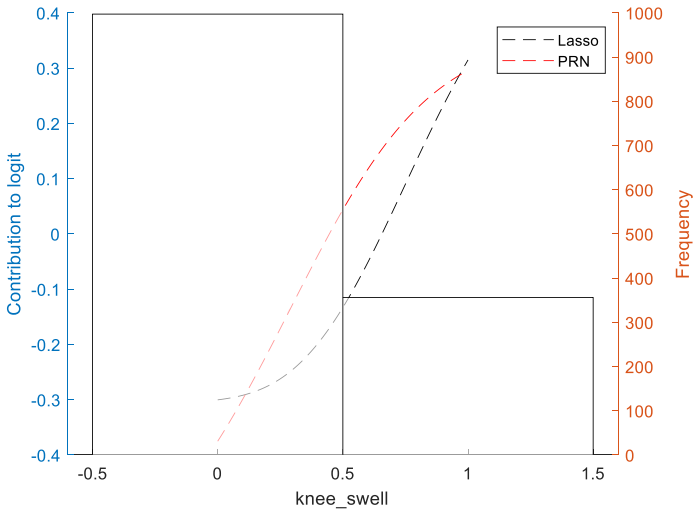


Figure 3-6: Partial response graphs for the variables in the final model as generated by the PRN. Graph A depicts age, B is BMI, C is presence of pain at initial investigation and D is knee swelling.

3.5. Discussion

The point of this chapter was to develop a diagnostic model that could be used as a tool to aid clinicians, and to determine the most appropriate model, through performance and interpretability, for classifying the KOA presence at baseline.

When developing a tool that will assist clinicians in performing diagnostics, it is important to compare to and establish the gold standard. When considering the 'gold standard' for diagnostic assessment of knee osteoarthritis radiography is widely accepted [100]. Whilst this under the form of x-ray and MRI are commonly used to diagnose KOA [101], the challenge with these methods is early recognition of the disease [102]. This is, in part, due to the cost of radiography and therefore individual practices trying to reduce costs by not sending large numbers of people for screening x-rays if they are not symptomatic. That

is where this model has the potential to be useful. It can help in primary care setting when an individual may not meet the standard symptoms of KOA and help signpost them to further diagnostic investigations, such as x-ray or MRI.

This broad range of commonly used models from conventional statistics and probabilistic ML shows broadly consistent performance in binary classification for diagnosis of Knee OA. They exploit the predictive power of a small set of covariates in each model, which are therefore the concluding set of predictive factors for diagnosis at first presentation.

From Table 3-1 CART analysis seems to be the worst performing model, but only slightly. However, the drop in performance can be traded off for simplicity and ease of interpretation that the model provides. The performance is important, but where decisions impact people it is imperative that results can be explained, and the CART model offers a high level of interpretability.

It remains to address the issue of model explanation. The rule sets are transparent by design. While the rules provide a filter for assigning patients to diagnostic categories, they do not provide a clear indication of the weight that each covariate has for the diagnostic inference made by the model for each individual patient.

This is provided only by logistic regression and can be conveniently expressed in the form of a nomogram, making the model easy to use and to interpret by clinicians. One such example in Figure 3-4 is generated using the LogR model. Moreover, the explicit weighting of covariates also provides a tool to 'diagnose the model', by correlating these weights against prior clinical expertise about the expected influence of each co-variate on the diagnostic outcome.

The most complex approach, MLP-ARD, does outperform the LogR model but is not interpretable or easily transferrable to a way that can be converted into simple rules that can be applied in clinical practice. The PRN-Lasso model was the most calibrated model but arguably not the most interpretable. The PRN-Lasso approach considers both univariate and bivariate features. The final PRN-Lasso model contained only four univariate features, but gained only a 3% performance improvement when considering the AUC when compared to the LogR model. Although this is a performance improvement, the LogR model is still the more interpretable model and is preferred method within clinical practice as the model can be displayed as a nomogram.

The small improvement from the slightly more complex approach provided by the PRN-Lasso would take more expertise in understanding the mechanics of the model than the LogR approach. The traditional method for modelling disease, LogR, has been the model of choice for many years as it replicates the human way of thinking. Each of the variables can be represented as switches that are either on or off, or contribute on a sliding scale to the overall outcome. This is useful for when trying to predict a disease outcome as these decisions are never black and white, but more often than not when based solely on symptoms come with a scale of how much the covariates contribute to the overall outcome. The nomogram for LogR is a useful way to transfer the model from a statistical and ML algorithm to a useable set of rules that can be followed in clinical practice.

Chapter 4: Survival Modelling

4.1. Introduction

The previous chapter describes a model that uses a selection of demographic and self-reported features with the aim of diagnosing KOA at the point of initial presentation. The model identifies which features are indicative of the presence of KOA. A useful next step is modelling the KOA free survival. This is defined as time from a disease free state at recruitment, to progression to clinical KOA. This instance of time to event analysis requires the key features of a start time, the point of recruitment, an end time, follow up for five years, and the event, the presence of KOA.

There are five stages of KOA according to the Kellgren-Lawrence (KL) scale [3]. These are differentiated between with the use of x-rays to determine the severity of the OA. Stage 0 is classed as no OA and Stage 4 is severe OA present in the joint. A clinician usually analyses and classifies images for diagnosis. By using both humans and machines there is the potential for more reliable diagnoses [4]. Stage 1 is the point at which disease changes are likely to begin but go unnoticed as they do not typically cause symptoms to the sufferer. Stage 2 is usually the point of diagnosis as this is where symptoms usually start to bother the subject. The advice often given at this stage of the disease is aimed at preventing progression. If behaviours can be modified prior to the onset of symptoms due to early interventions, then the burden of OA is likely to be reduced.

The progression through the disease stages does not follow a linear pattern. Loss of cartilage is a primary factor in OA development. There are three main ways that loss of cartilage can occur: slow and progressive taking decades, rapid deterioration over 12-24 months or periods of time that are fluctuating between stunted and rapid progression [103].

With progression a key point of interest, survival modelling is the approach used to look at this. Survival analysis approaches can be used to make inferences about time to onset of a disease [104]. Survival analysis also has many other uses where time-to-event is measured, such as mechanical failure and average employee turnover [105]. It is also known as failure-time analysis [106].

The two primary methods to analyse survival curves are the Kaplan-Meier method and Cox proportional hazards regression [107], [108]. These methods contribute importantly to medical statistics, often used to define the prognostic features for disease onset or development [109]. These approaches have been used in many medical applications such

as oncology for decision support [110]. Kaplan-Meier is an observational approach that can only be applied to existing data, whereas Cox regression models can be applied to prospective data to generate predictions.

In the medical arena, survival analysis is important and versatile as it offers the chance to analyse any number of event outcomes, such as recovery, time to clinical intervention, disease onset or death. It is also applied to a wide variety of diseases. The approach is intended to make a decision, usually about therapy, at the point of recruitment. This allows a glimpse forward to inform decision made at the point of recruitment. Here, for example, a possible decision could be to either prescribe possible disease modifying medication, or give advice based on a projected outcome for disease progression.

Cancer is a particular disease that is modelled frequently with survival methods [111]. The power of using approaches related to survival is that the results can suggest one treatment type, or set of initial conditions that lend itself to an outcome that is more likely than another given a set of initial conditions. For example, a study in Japan spanning 28 years using 173,378 patients with hepatocellular carcinoma suggests that the five-year overall survival rate for someone diagnosed between 2001-2005 is 58.4% following resections, whereas following ablation survival is 47.6% for the same window [112].

Other uses of survival analysis include looking at time to failure of kidney grafts [113] and analysing the failure rate of dental implants in the first year in diabetes patients [114].

When looking at the area of knee osteoarthritis, survival modelling has predominately focused on progression from an arthritic state to joint replacement. One example of this is looking at the Importance of cartilage defects in older adults in relation to progression to knee replacement [115]. A similar study investigated the incorporation of radiographs when predicting the likelihood of total knee replacement within 9 years and the final Kellgren-Lawrence grade [116].

Some studies focus on the likelihood of developing KOA following certain treatment courses. For example, one such study looked at the risk of requiring knee replacement surgery following treatment with intra-articular corticosteroid injections [117]. A similar study found that the use of intra-articular corticosteroid injection increases the risk of KOA progression [118].

An approach using an increase in joint space narrowing as an outcome in survival modelling has also been investigated in subjects with known symptomatic OA [119]. This

study found that once radiographic changes were visible then the risk of progression in OA was significant.

There has even been a study looking at the relationship between depressive symptoms and OA. This analysis found that in people with depressive symptoms their risk of developing OA increased [120]. In a similar way to the depression study, there has been a study looking at the perceived quality of life in different populations relating to OA, those with higher prevalence of pain medication and those who have previously undergone knee replacement [121]. This study found that those with knee OA and those taking pharmaceutical interventions suffer with a lower perceived health related quality of life.

There is currently a gap looking at covariates in a population with no KOA, and how these covariates influence the risk of onset. This work looks at the same variable set used in the diagnostic model to determine whether they can also be useful in predicting the risk of progression from a disease free state to one with clinical KOA.

Survival modelling has huge potential in relation to KOA. The survival analysis model, if used in clinical settings, would help to be a tool useful for patient education by clearly showing time to progression before behaviour modification, along with giving timelines for clinicians for development of estimated treatment plans to help optimise disease management.

Chapter aims

- Investigate both the 7-year and 5-year survival using Kaplan-Meier curves and Cox regression on the useable cohort from the OAI data.
- Determine which features are significant in the development of KOA.
- Establish what features contribute to an increased risk of developing KOA and which are attributed to a lower risk of developing KOA.

4.2. Specifics of the data

The data used in this analysis is from the Osteoarthritis Initiative (OAI) [62]. The full explanation of the data is in Chapter 2.

As mentioned previously, certain variables have been categorised and although they lose some information, this approach can provide more use in a patient facing setting. For the survival modelling specifically BMI was categorised into three groups: BMI below 25,

BMI 25-29.9 and BMI 30+. When looking at how best to categorise the BMI variable different approaches were considered.

The ideal BMI is in the range 18.5 to 24.9, with anything above or below this likely to have health implications [122]. For this dataset, there were very few cases of people with a BMI less than 18.5 ($n = 7$). The NHS define the BMI groups as:

Below 18.5	– Underweight,
Between 18.5 – 24.9	– Healthy weight range
Between 25 – 29.9	– Overweight range
Between 30 – 39.9	– Obese range
Above 40	– Clinically Obese

The main question was to consider if two categories, such as those used in the diagnostic modelling, would be most beneficial or if incorporating a third category would offer additional insight.

After the analysis looking at two and three BMI splits, the survival analysis modelling will have BMI in 3 categories: BMI less than 25, Overweight (25-29.9) and obese (30+). The p-value on the three groups is $p < 0.0001$, the Kaplan-Meier curve can be shown in Figure 4-1. When considering a p-value of 0.05 to be statistically significant anything that meets this or is smaller is also significant. The p-value is the probability of rejecting the null hypothesis, given that the null hypothesis is true. In this instance, the null hypothesis is that there is no difference between the BMI groups. As the p-value is less than 0.0001 the null hypothesis is unlikely, therefore the difference is statistically significant.

Even though both BMI 25/25+ and BMI 30/30+ are statistically significant, the 3 group BMI split offers the chance to be more informative, therefore offering more insight into the cohort. An additional benefit from having BMI in one of three categories is that the subject can see the impact of simply reducing their BMI from one category to another.

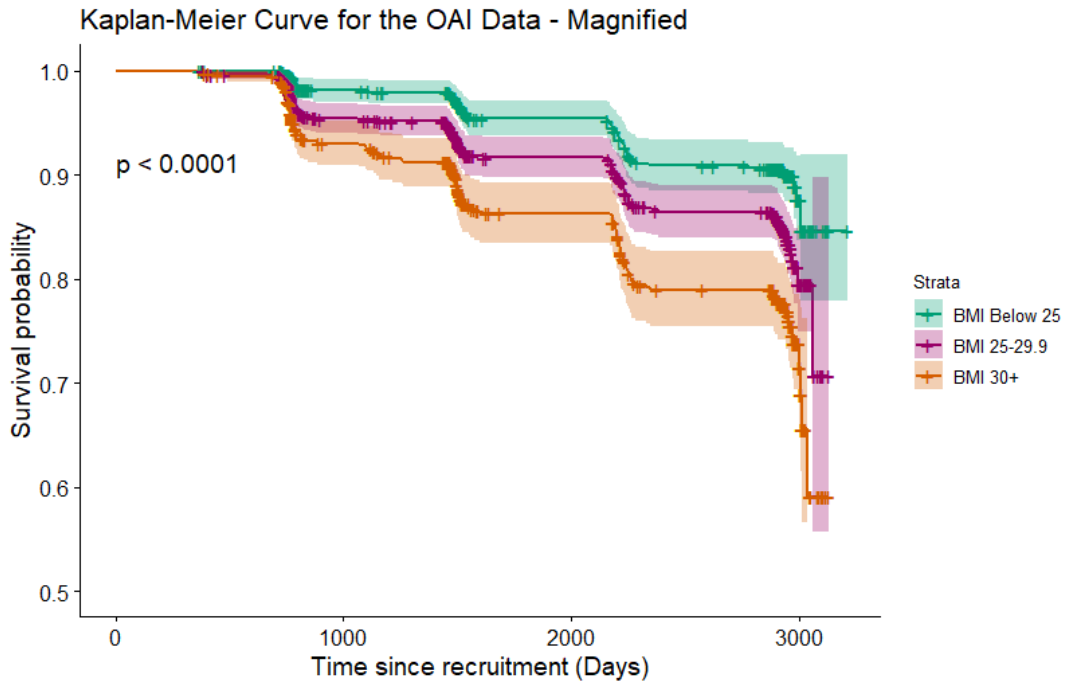


Figure 4-1: Unadjusted Kaplan-Meier with the BMI separated in the bins of below 25, between 25 & 30 and 30 & above. The separation of these groups is significant with a p-value smaller than 0.0001.

4.3. Cohort Definition

For survival analysis, the cohort of subjects needs to be defined.

The usable cohort meeting the criteria for survival analysis is comprised of 2136 subjects. The cohort of data has low prevalence of cases developing to KL2+. There are 117 subjects that go on to develop KOA during the follow up window, and there are 2019 subjects who did not develop KOA in the follow-up period, shown in Figure 4-3. Figure 4-3 shows the two bar charts overlapped, to illustrate the extreme difference in cases vs censored data. These subjects are right censored. This is where the subject either does not develop the disease in the follow up period, or at some point the subject is lost to follow up, but at the time of the last clinical visit had not yet developed the disease. The sample is made up of 94.5% censored data. Figure 4-2 shows the way that the data has been pre-processed for the usable cohort suitable for the survival analysis.

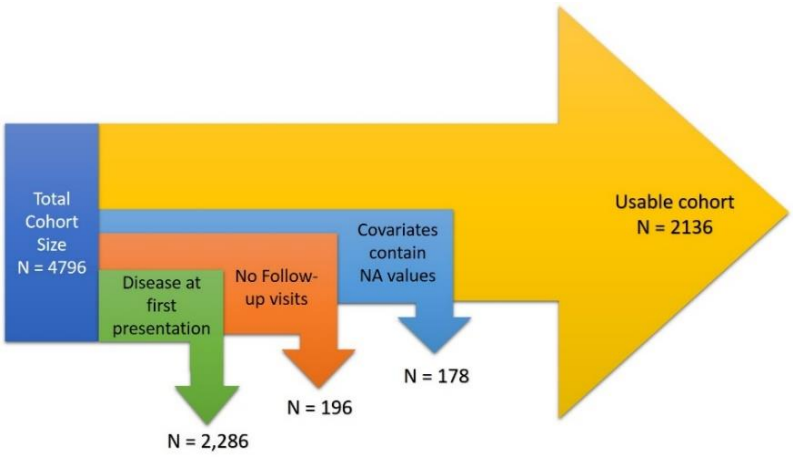


Figure 4-2: A Sankey diagram to visually display the data pre-processing.

Figure 4-3 shows how cases are split, between event (developed to OA, outcome 2) and censorship (subjects who are either lost to follow up or do not develop OA by the end of the study window, outcome 1). Because of the low prevalence of cases in the data, along with the increased variability in covariates as time passed increases, different cohorts are considered.

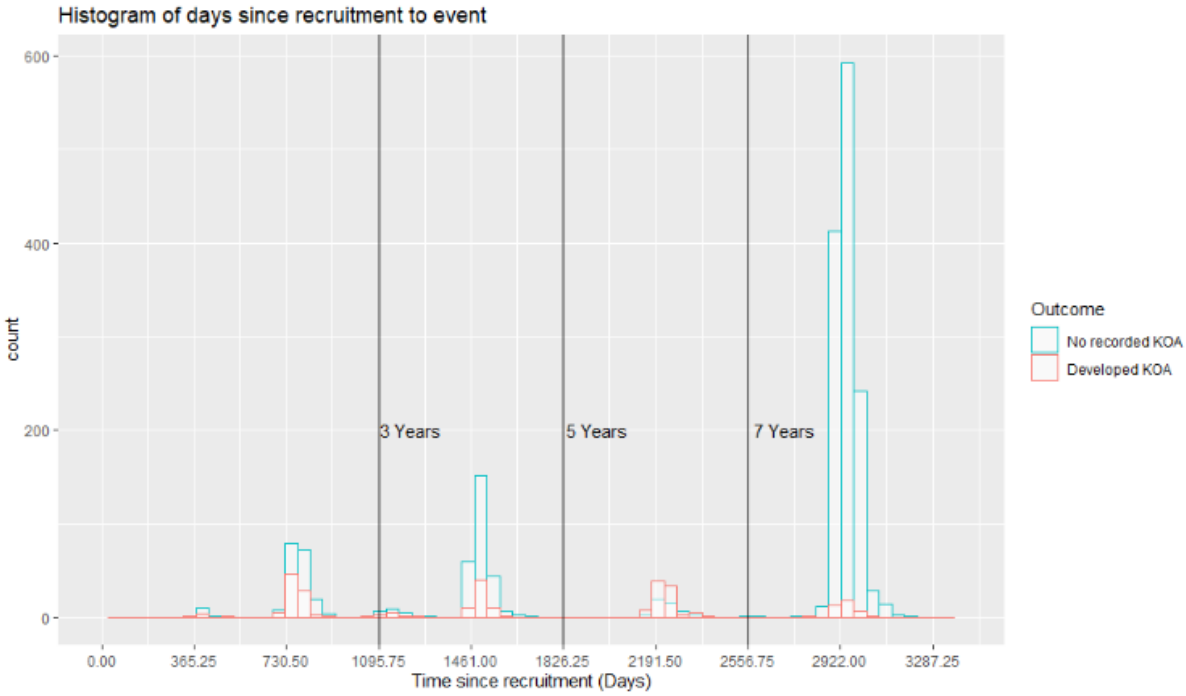


Figure 4-3: Bar plot showing the days since recruitment to event - either OA recorded or censorship.

The rationale for selecting the given cohorts will be detailed in section 4.3.1.

4.3.1. Cohort Selection

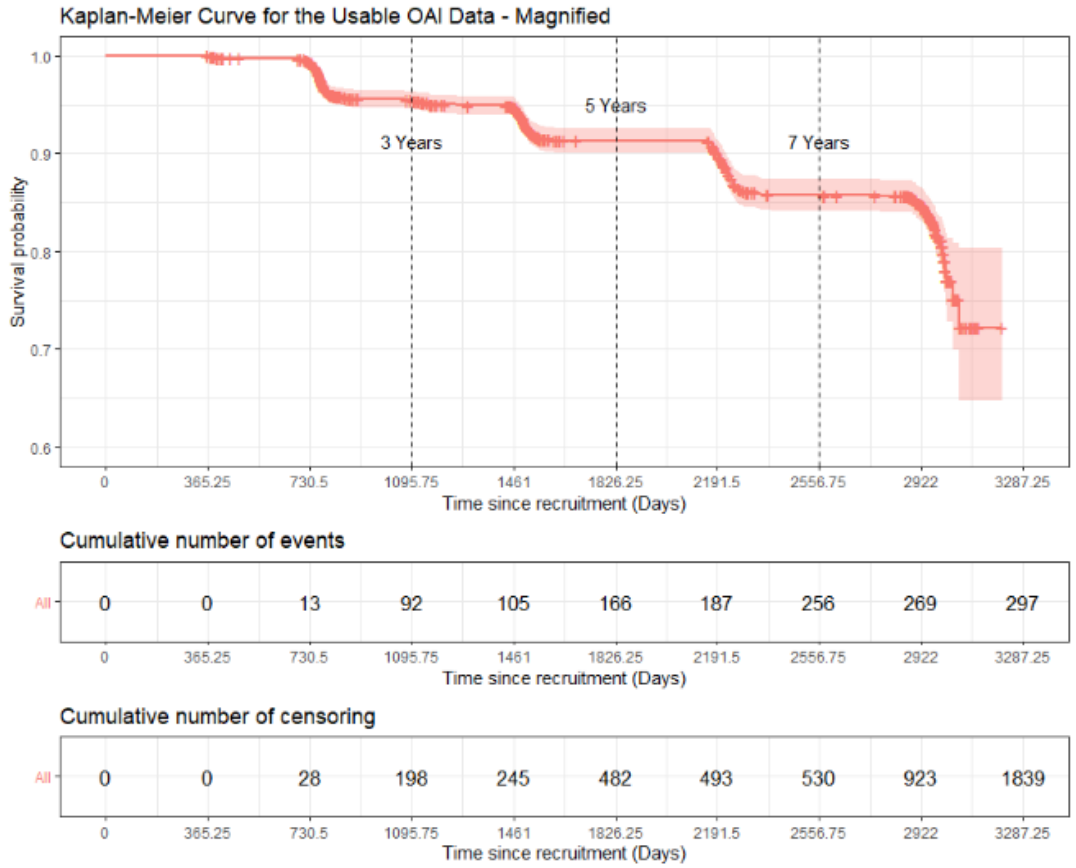


Figure 4-4: Kaplan-Meier curve for the whole population, with a table detailing, at yearly intervals, the cumulative number of events and censoring.

It can be seen in Figure 4-4 that the number of events in the whole cohort is 297, where there is a very large (1839) number of censored cases. The vertical lines on the KM curve shows how the cohorts are divided for analysis in this analysis. The following analysis considers the seven and five-year cohorts. The three-year cohort is not considered as the window from initial to follow up does not have a high amount of follow up. For the three year analysis to be more suitable for investigation, the subjects would have required follow up visits at 3 or 6 month intervals to ensure a more complete sample to model with.

4.3.2.1. Seven-year cohort

The seven-year cohort is made up of 2,095 subjects, shown in Figure 4-5. There are 255 instances where a subject goes on to develop OA in the follow up window, and 1,840 where the subject was censored. The total cohort differs by 41 subjects as they develop KOA after the 7-year cut off point.

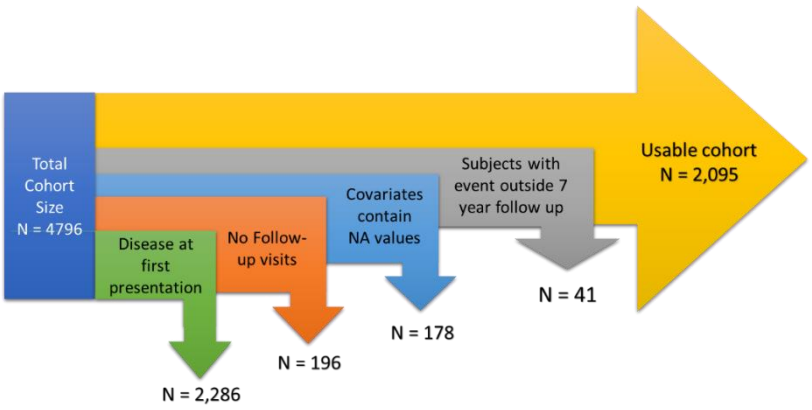


Figure 4-5: A Sankey diagram to highlight the number of subjects excluded from the 7-year analysis, showing how the data pre-processing occurred and highlighting the final usable cohort of 2,095 people.

The histogram in Figure 4-3 shows the split in cases between OA and Non-OA. This is a visual representation of the way the events occur in the timeframe of interest. For example, most events of both types occur around the four-year mark.

The seven-year cohort is split into train/test splits. The justification for this is given in Chapter 2, section 2.7. The training set consists of 1047 subjects and the test sample has 1048 subjects. Table 4-1 describes how the data is split into the training and test splits.

Table 4-1: Description of the way that the training and test samples are made up.

	Total Data Size	Censored	Disease development
Training Set	1047	908	139
Test Set	1048	932	116

4.3.2.2. Five-year cohort

The five-year cohort is made up of 2,005 subjects, shown in Figure 4-6. There are 166 instances where a subject goes on to develop OA in the follow up window, and 1,839 where the subject was censored.

The rationale for selecting the five-year cohort will be detailed in section 4.5.2.

The five-year cohort is split into train/test splits. The training set is made up of 1002 samples and the test sample has 1003 subjects. Table 4-2 describes how the data is split in the training and test splits.

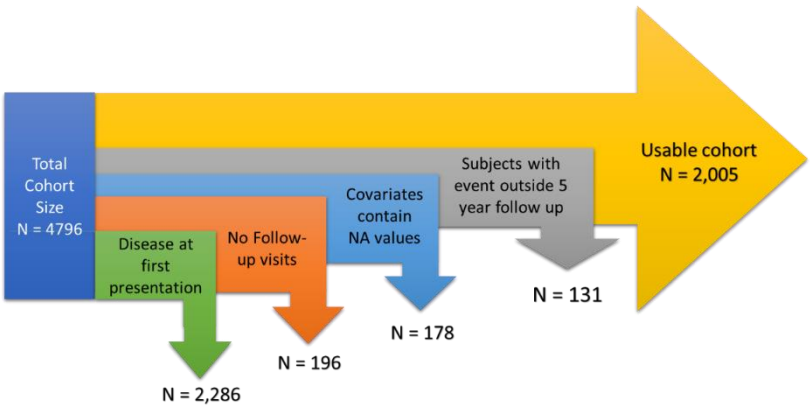


Figure 4-6: A Sankey diagram to highlight the number of subjects excluded from the 5-year analysis, showing how the data pre-processing occurred and highlighting the final usable cohort of 2005 people.

Table 4-2: Description of the way that the training and test samples are made up.

	Total Data Size	Censored	Disease development
Training Set	1002	913	89
Test Set	1003	926	77

4.4. Methods

4.4.1. Survival and Hazard Functions

Two related functions are used to describe survival data: the survival probability and the hazard function [106].

The survival probability, often referred to as the survivor function $S(t)$, is the probability that an individual survives from the time origin to a specified future time t.

The hazard function, denoted by $h(t)$, is the instantaneous rate of occurrence for an individual who experiences the event.

4.4.2. Kaplan-Meier

Kaplan-Meier (KM) is an empirical survival function that describes a patient's survival over time [108]. The KM estimator is a non-parametric statistic that allows us to estimate the survival function [123]. This non-parametric approach is not based on the assumption that there is an underlying probability distribution. This is useful as survival data often has a skewed distribution. The Kaplan-Meier formula is shown in Equation 4-1.

The KM statistic gives the probability that an individual patient will survive beyond a particular time t [124]. At $t = 0$, the KM estimator is 1 and with t going to infinity, the

estimator goes to 0. In theory, with infinitely large data, and t measured to the second, the function of t versus the survival probability is smooth.

It is further based on the theory that the likelihood of surviving past a certain time point t is equal to the product of the observed survival rates until time point t. More precisely, $S(t)$, the survival probability at time t is given by

$$S(t) = p1 \times p2 \times \ldots \times pt$$

Equation 4-1: The survival probability formula for Kaplan-Meier calculations.

With $p1$ ~ the proportion of all patients surviving past the first time,

$p2$ ~ the proportion of patients surviving past the second time point

... ~ this is the proportion of surviving patients until the time point t is reached.

It is important to note that starting at $p2$ up to pt the only patients that are considered are those that survived past the previous time point when calculating the proportions for each next time point; therefore $p2, p3, \ldots, pt$ are all proportions that are conditional on the previous proportions.

This can also be calculated by:

$$S(t_i) = S(t_{i-1})\left(1 - \frac{d_i}{n_i}\right)$$

Equation 4-2: The estimated probability formula of being at risk of disease at time t_{i-1}.

Where $S(t_{i-1})$ ~ the probability of being at risk at t_{i-1},

n_i ~ the number of patients at risk just before t_i,

d_i ~ the number of events at time t_i,

$t_0 = 0, \quad S(0) = 1$

And, $\frac{d_i}{n_i}$ is the hazard, the risk of the event of interest.

The estimated probability $S(t)$ is a step function that changes value only at the time of each event, Equation 4-2. It is also possible to compute confidence intervals for the survival probability.

The KM survival curve, a plot of the KM survival probability against time, provides a useful summary of the data that can be used to estimate measures such as median survival time [106].

4.4.3. Cox Regression

The purpose of the Cox model is to simultaneously evaluate the effect of several factors on survival [107]. It allows us to study how specified factors impact the rate of a particular event happening at a particular point in time. This rate is commonly referred to as the hazard rate. Predictor variables are referred to as covariates.

The Cox model is expressed by the hazard function denoted by $h(t)$ [125], shown in Equation 4-3. The hazard function can be interpreted as the risk of dying, or in this case, contracting KOA, at time t. It can be estimated as:

$$h(x_1, x_2, \dots, x_p, t) = h_0(t) \times exp(b_1 x_1 + b_2 x_2 + \cdots + b_p x_p)$$

Equation 4-3: The hazard function for contracting KOA at time t.

Where t ~ survival time

$h(t)$ ~ the hazard function determined by a set of p covariates $(x_1, x_2, \dots, x_p)$. The t shows that the hazard may vary over time.

$(b_1, b_2, \dots, b_p)$ ~ coefficients that measure the impact of the covariates

h_0 ~ baseline hazard. It corresponds to the value of the hazard if all the x_i are equal to zero $(e^0 = 1)$.

The hazard function, Equation 4-3, factorises the hazards by separating the time dependency from the covariate dependency. The baseline hazard, h_0, is calculated using a reference population, where all risks are at the baseline state. The hazard for everyone else is then modelled as proportional to the baseline hazard.

The Cox model can be written as a multiple linear regression of the logarithm of the hazard on the variables x_i, with the baseline hazard being an 'intercept' term that varies with time [126]. The quantities $exp(b_i)$ are called hazard ratios (HR). A value of b_i greater than zero, or equivalently a hazard ratio greater than one, indicates that as the value of the i^{th} covariate increases, the event hazard increases and thus the length of survival decreases [127]. A hazard ratio above one indicates a covariate that is positively

associated with the event probability and therefore negatively associated with length of survival.

Another function that is useful in the context of survival analysis is the hazard function $h(t)$. It describes the spontaneous rate of occurrence of the event, h, if the subject survived up to that particular time point, t [128]. It is slightly more difficult to illustrate than the KM estimator because it measures the instantaneous risk of the event. Nevertheless, you need the hazard function to consider covariates when you compare survival of patient groups. Covariates, also called explanatory or independent variables in regression analysis, are variables that are possibly predictive of an outcome or that you may want to adjust for to account for interactions between variables [129].

Whereas the log-rank test compares two KM survival curves, which might be derived from splitting a patient population into treatment subgroups, Cox proportional hazards models are derived from the underlying baseline hazard functions of the patient populations in question and an arbitrary number of dichotomized covariates [129]. Again, it does not assume an underlying probability distribution but it assumes that the hazards of the patient groups you compare are constant over time. That is why it is called 'proportional hazards model' [130].

The Cox proportional-hazards model is essentially a regression model commonly used statistical in medical research for investigating the association between the survival time of patients and one or more predictor variables [125]. Kaplan-Meier curves and log-rank tests are useful only when the predictor variable is categorical. They do not work easily for quantitative predictors such as gene expression, weight, or age.

An alternative for the KM approach is the Cox proportional Hazards regression analysis [123]. This works for both quantitative and categorical predictor variables. Also, the Cox regression model extends survival analysis methods to assess simultaneously the effect of several risk factors on survival time.

4.4.4. Akaike Information Criterion

The Akaike Information Criteria (AIC) is a predictor of the relative quality of a model for a given dataset [131]. AIC is a concept that is founded on information theory, and is used in model development. In this thesis, the AIC is used in the stepwise feature selection to help construct the model. The formula is shown in Equation 4-4.

$$AIC = 2k - 2\ln(L)$$

Equation 4-4: The Akaike Information Criterion formula.

Where k is the number of degrees of freedom used,

L is the partial likelihood

$$L = \prod_{i=1}^{n} \left\{ \frac{\exp(\beta' x_i)}{\sum_{l \in R(x_i)} \exp(\beta' x_l)} \right\}^{\delta_i}$$

Where $R(t_i)$ is the risk set at time t_i,

n is the observed survival times from time $t_1, t_2, \ldots, t_n$,

δ_i is the event indicator, this will be zero when t_i is censored,

x_i is the vector of covariates for the individual whose events occurs at the i^{th} ordered time t_i,

The summation is the sum of the values of the $exp(\beta' x)$over all individuals who are at risk at time t_i.

4.4.5. Stepwise Feature Selection

Stepwise feature selection uses the AIC as a measure to determine which terms combine to form a model with a better fit to the data. A lower score can indicate a more frugal model, with fewer features, when compared to one with a higher AIC value. By using the AIC to remove terms deemed to not be beneficial to the model, only features that add value to the model are used. This helps to limit the number of features in the model. The stepwise approach uses a step-by-step iterative approach to construct a regression model. The act of stepwise regression will either add or remove potential explanatory variables in succession and testing for statistical significance after each iteration of the feature selection.

As with any approach, stepwise feature selection has advantages and disadvantages that need to be considered before use in modelling. One advantage in using stepwise is that it improves the models ability to generalise. This happens as stepwise inherently reduces the number of predictors in the model, therefore improving the out-of-sample accuracy. The stepwise feature selection also offers the advantage of a simple model with easy interpretation due to the reduced number of variables. By utilising an automatic algorithm

when selecting variables helps to eliminate bias that can be present when relying solely on expert opinion. In the case presented in this thesis a combination of expert based selection followed by stepwise has been used to select variables for the final models. This step adds a layer of objectivity in selection of the features that are to be included within the models. However, as with any method there are limitations. One such limitation is that stepwise feature selection, both forward and backward, does not consider all potential combinations of predictors. This in itself can have a computational advantage when there are a large number of possible combinations to test but it does mean that there is no guarantee that the final combination of variables is in fact the best combination possible. Another disadvantage is that using stepwise feature selection can produce an unstable selection of variables. One way to counter this is to reduce the original variable pool through expert input or surveying existing literature. As this step was taken in this thesis, the variable selection is stable.

Forward stepwise feature selection initially starts with no variables in the model and works by adding each new variable incrementally and at each point the AIC is calculated. This value is then compared with the previous model and if the AIC is lower, another variable is added and the process is then repeated until the AIC no longer decreases.

The process works similarly for backward stepwise feature selection, only the model uses AIC calculates and removes a variable, then recalculated the AIC and compares the two. This is repeated whilst the AIC value is decreasing. Backward stepwise starts with a full model and removes features one at a time to test its importance, determining if the removed variables are statistically significance.

4.4.6. Test for Proportional Hazards Assumptions - Schoenfeld Residuals

There is a consistent downside to other residuals such as Cox-Snell and the Martingale residual. This is that they rely heavily on observed survival time and therefore require an estimate of the cumulative hazard function. The Schoenfeld residual developed in 1982 overcomes these [132].

This residual produces a value for each explanatory variable in the fitted Cox model. The Schoenfeld residual is described in Equation 4-5 and Equation 4-6.

The i^{th} Schoenfeld for X_j the j^{th} explanatory variable in the model is given by

$$r_{pji} = \delta_i\{x_{ji} - \hat{a}_{ji}\}$$

Equation 4-5: The Schoenfeld Residual equation

Where x_{ji} is the j^{th} variable $j = 1, 2, \dots, p$ for the i^{th} individual in the study.

$$\hat{a}_{ji} = \frac{\sum_{i\epsilon\mathbb{R}(t_i)} x_{ji}\, exp(\widehat{\beta}' x_i)}{\sum_{i\epsilon\mathbb{R}(t_i)} exp\, \widehat{\beta}' x_i}$$

Equation 4-6: Schoenfeld residuals formula to test if the proportional hazards are independent of time.

Where $\mathbb{R}(t_i)$ is the set of all individuals at risk at time t_i, $\widehat{\beta}' x_i = \widehat{\beta_1}' x_{1i} + \widehat{\beta_2}' x_{i2} + \cdots + \widehat{\beta_p}' x_{pi}$ is the value of the fitted component, linear predictor, of the model for that individual.

Hazards are said to be proportional if the ratios of hazards are independent of time. If one or more of the variables are time dependant or vary over time, then the assumption of proportional hazards is violated.

4.4.7. Stratification of Risk Groups

In order to create cohorts and profile the risk of an individual developing KOA the first step is to stratify the population. To do this, the point at which the cut is between groups needs to be identified.

The Cox model produces predictions of the risk score for the training dataset. These are plotted into a histogram, displaying the distribution of risk scores, showing a Chi^2 distribution. The predictions are shown in bins of 0.25 from 0 to the maximum value calculated wising the original Cox model.

At each bin interval for the risk score the values above the cut point are assigned to cohort 1 and those below are assigned to cohort 2. Using the cohorts, a Cox model is fitted to the data and two baseline hazards are fit to the strata. The next step is to test the strata to see if there is a difference between the curves with a log rank test. This test produces a Chi^2 statistic, which is used to determine the p-value relating to the cohort stratification. This is repeated for each bin interval of the risk score, storing the p-values at each iteration. To identify the risk score that relates to the optimal cut point for the cohorts the minimum p-value is determined and the corresponding risk score is selected. This is then the score used to split the population into two cohorts.

4.5. Study Design

Initially, after looking at the Kaplan-Meier curve in Figure 4-4 it was decided that the 7-year cohort would be used. This is due to the large amount of censoring that occurs after the seven-year point. After the seven-year point there is a high level of uncertainty in the data due to the level of censorship. The seven-year cohort would look for indicators to development of clinical KOA.

It was only after concluding the seven-year cohort study that it was deemed necessary to investigate using the five-year cohort, as the predictions on the test data for the survival after the 5-year mark vary greatly from the training data. These fall outside of the confidence interval and therefore are not as reliable, and resulted in a change in dataset timeframe.

4.5.1. Consideration of 7 - year cohort

The analysis in this section of work will consider the seven-year cohort. A seven year follow up is a long time when considering some covariates, such as pain when walking, taking medication and falling, as although the methods in this report consider static variables and following the assumption that these variables remain constant, in practice this is not always the case.

A seven-year follow-up for trying to identify if a disease occurs with the outlook to patient education in order to modify behaviour gives a meaningful window to the future. A seven-year outlook is something that a person can still look to without it being a very long follow-up, such as the study length of nearly 9 years.

Figure 4-3 illustrates a histogram of the total cohort, showing the event type as time progresses. The Kaplan-Meier curve in Figure 4-4 shows the whole cohort and the events in the timeline from study enrolment to the end of follow up. Table 4-3 shows the differences in the cohorts between the total cohort, 7-year and 5-year cohorts. Initially, for maximisation of information, the 7-year cohort is used.

Table 4-3: Description of how data is split between the development of OA and censored for each cohort of interest.

Cohort	Develop OA	Censored	Total Data Size
Total Cohort	297	1839	2136
7 Year Cohort	255	1840	2095
5 Year Cohort	166	1839	2005

4.5.2. Consideration of 5 year Cohort

Following on from the stratification in the analysis for the 7-year cohort, the predictions on the test dataset vary greatly from the actual training values. Upon further investigation, the curves are only similar up to the 5-year mark, where they begin to diverge quite significantly.

The total loss in information between the five and seven year cohorts is not that much, as shown in Table 4-3, and the five-year cohort still holds clinical purpose, which makes it useful for analysis. In order for an intervention to work, there needs to be some level of subject acceptability, referring to the suitability of the intervention to both those delivering and receiving the care [133]. When trying to implement change into the way a subject behaves based on a potential outcome then a nearer end-point can be seen as more advantageous. A similar approach is used when trying to encourage people to quit smoking, displaying relatively short time steps into the future, given with the associated benefit [134]. One such example is that 48 hours after quitting smoking a person's taste and smell receptors begin to heal [135].

4.5.3. Plan of work for analysis

The work that follows will be the application of Cox regression to different variable subsets of the data in each case, the seven and five-year datasets.

- Cox regression on whole variable cohort
- Cox regression on Backward stepwise variable cohort
- Cox regression on Forward stepwise variable cohort
    - If the forward and backward models are not consistent, perform a Cox regression on stepwise overlap cohort. This means that if the forward and backward features are not consistent, a model comprised of only the consistent features will be considered in a Cox regression model.
- Cox regression with stratification on the feature selected model

4.6. Results from Analysis

The results in this chapter will be split into two sections: the results from the seven-year cohort, and the results from the five-year cohort.

4.6.1. Results from seven - year cohort

The stages from the seven-year cohort study were to use all of the available covariates,

and then to use a subset of these selected by stepwise feature selection, both backward and forward.

4.6.2.1. Kaplan Meier Curves

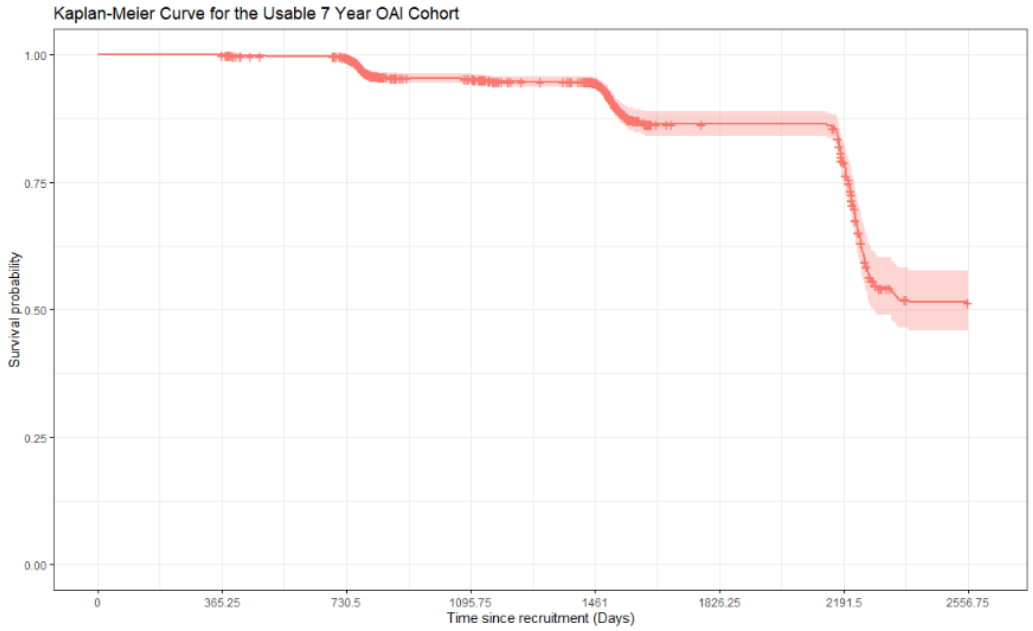


Figure 4-7: A Kaplan-Meier curve for the data of the 7-year cohort.

Figure 4-7 is the Kaplan-Meier curve for the usable 7-year OAI cohort. The steps in this plot are artefacts in the data, likely due to the follow-up design in the study, in which a certain window of time passes between follow-up assessments. Therefore, the true survival curve would likely be a smooth curve between the steps, however due to the follow-up that is not the case; therefore, any models generated with the OAI data would require further tests to determine suitability for use on different data.

An initial exploration into discrete time survival analysis is covered in section 4.7.

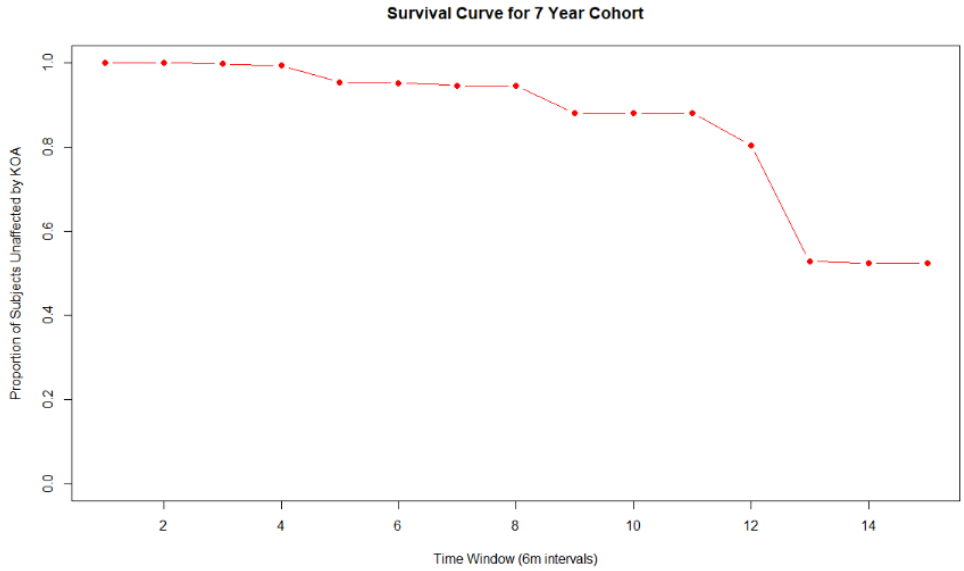


Figure 4-8: The curve developed by calculating the survival information in the life table.

The curve shows the actual curve depicting way the overall survival probability changes over time as generated in R by using the whole population of the study to assess the global survival when compared to the same cohort where the measures were calculated manually in the life table.

A step used to get to know the data was to plot the Kaplan-Meier curve and calculate the survival probability manually, shown in Figure 4-8 and Table 4-4. The reasoning behind this was to determine how the plots in R were generated, and the survival calculated. This step was necessary in confirming the plots generated in R, such as the plot in Figure 4-7, were correct and showed the survival probability as a true reflection of the data.

Table 4-4: Life table showing the calculation for the 6-month intervals of KOA survival over a seven-year period.

Day Intervals	№ okay at start of interval a_i	Develop disease during interval d_i	№ censored (lost to follow up) during interval c_i	№ of persons at risk $n_i = a_i - \frac{c_i}{2}$	Risk of onset during interval $r_i = \frac{d_i}{n_i}$	Chance of surviving time interval $s_i = 1 - r_i$	Cumulative chance of surviving from start of investigation $S(i) = S(i-1) \times s_i$
$0 \leq x < 182.625$	2095	0	0	2095	0	1	1
$182.625 \leq x < 365.25$	2095	0	0	2095	0	1	1
$365.25 \leq x < 547.875$	2095	6	13	2088.5	0.00287	0.99713	0.99713
$547.875 \leq x < 730.5$	2076	7	20	2066	0.00339	0.99661	0.99375
$730.5 \leq x < 913.125$	2049	77	298	1900	0.04053	0.95947	0.95347
$913.125 \leq x < 1095.75$	1674	2	4	1672	0.00120	0.99880	0.95233
$1095.75 \leq x < 1278.375$	1668	11	29	1653.5	0.00665	0.99335	0.94600
$1278.375 \leq x < 1461$	1628	2	133	1561.5	0.00128	0.99872	0.94479
$1461 \leq x < 1643.625$	1493	61	1183	901.5	0.06767	0.93233	0.88086
$1643.625 \leq x < 1826.25$	249	0	3	247.5	0	1	0.88086
$1826.25 \leq x < 2008.875$	246	0	0	246	0	1	0.88086
$2008.875 \leq x < 2191.5$	246	21	11	240.5	0.08732	0.91268	0.80394
$2191.5 \leq x < 2374.125$	214	67	36	196	0.34184	0.65816	0.52912
$2374.125 \leq x < 2556.75$	111	1	1	110.5	0.00905	0.99095	0.52433
$2556.75 \leq x < 2739.375$	109	0	109	54.5	0	1	0.52433

Figure 4-9 shows the KM curve for the 7-year cohort. The lines represent the full cohort, training, and test data. This shows that the training and test samples are a reflection of the whole cohort.

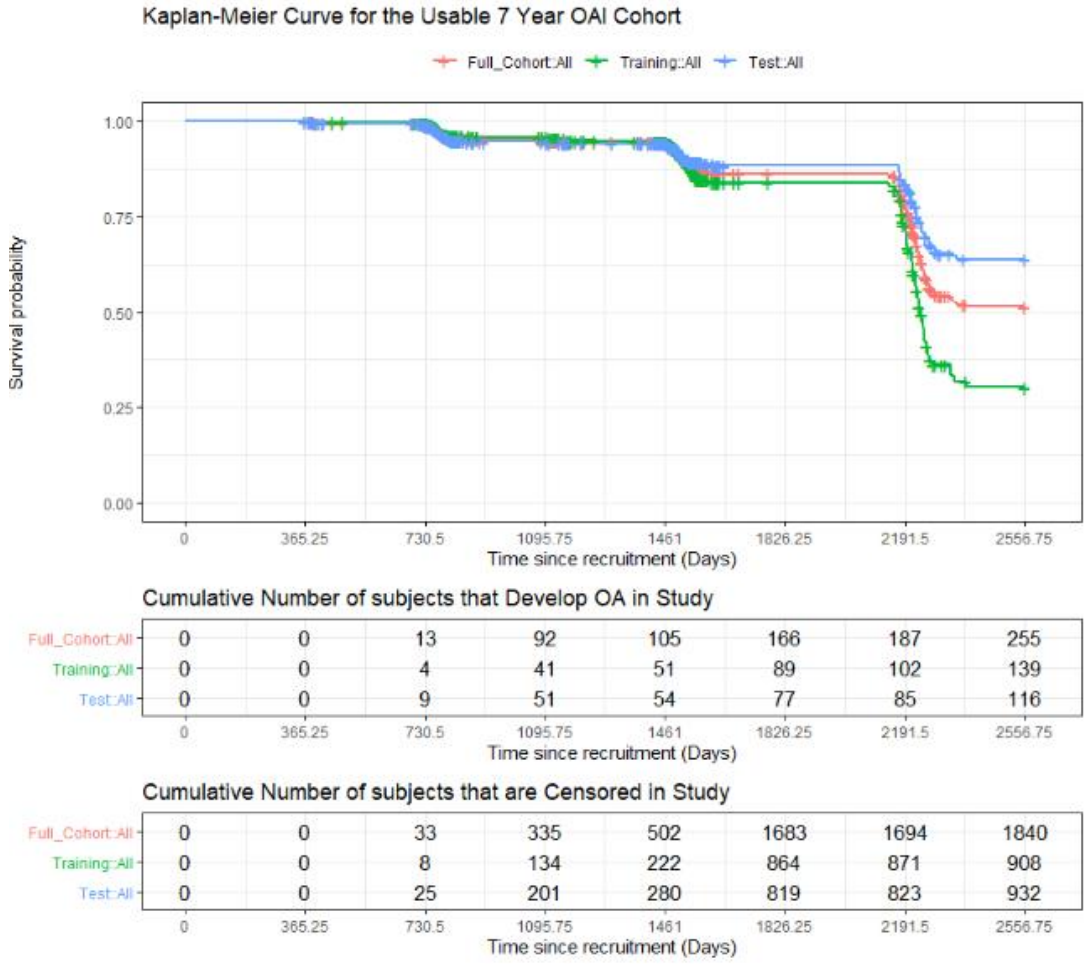


Figure 4-9: KM curve stratified by sample. The red depicts the total sample of the 2095 subjects in the study. The green shows the training sample, and the blue shows the test sample. The tables below illustrate the way in which the data is split between the samples.

4.6.2.2. Cox Regression

4.6.2.2.1. Univariate

The univariate model was conducted as a way to do exploratory data analysis to determine which, if any, are alone significant in development of KOA. This approach was used to identify a link between the variables and the outcome, development of KOA, and highlight features to consider in the multivariate models, establishing which, if any, are still significant when considered with others. Based on the results shown in Figure 4-10, the following variables are significant to development; Ever have a knee injury (EV.INJ), maxWOMAC, baseline symptoms at first assessment (B.LINE_SYMP), minPAIN, minSYMP, previous OA diagnosis (PREV.OA) and BMI. We know from the literature that features such as gender, genetic disposition, BMI and history of injury are all factors that contribute to the onset of KOA [77]. Therefore, these results make sense in practical

terms.

maxWOMAC, minSYMP and minPAIN are variables based on a questionnaire given to the patient. maxWOMAC is the maximum score calculated using the Western Ontario and McMaster Universities Arthritis Index. It is used in the evaluation of hip and knee OA, and considers features relating to pain, stiffness, and physical function. The WOMAC score can have a value from 0 to 96, with WOMAC a high score relating to a more severe impact. minSYMP and minPAIN are self-perceived pain and symptom scores calculated using the KOOS score. KOOS is the knee injury and osteoarthritis outcome score which uses patients own opinions to evaluate how they perceive their condition. KOOS uses five subscales but this analysis focuses only on symptoms and pains. KOOS scores of 0 indicate extreme problems with higher scores indicating fewer problems. KOOS is mainly used where knee injury can result in post-traumatic OA.

The reasoning for inclusion of both WOMAC and KOOS measures are that although WOMAC considers all KOA. KOOS is aimed at KOA following traumatic injury and by including both there is a better chance that a subjects symptoms will be accounted for and considered in the model.

	beta	HR (95% CI for HR)	wald.test	p.value
AGE	-9.1e-05	1 (0.98-1)	0	0.99
FAMILY_H	-0.24	0.79 (0.5-1.2)	1	0.31
SYMPTOMS	0.52	1.7 (0.93-3)	2.9	0.087
EV.INJ	0.71	2 (1.4-2.8)	17	4.1e-05
HAND.OA	-0.074	0.93 (0.58-1.5)	0.1	0.75
maxWOMAC	0.015	1 (1-1)	8.7	0.0032
HIST.FALL	0.28	1.3 (0.95-1.9)	2.8	0.097
KNEE.SURGERY	0.21	1.2 (0.81-1.9)	0.92	0.34
PAIN.MEDS	0.11	1.1 (0.64-1.9)	0.15	0.7
B.LINE_SYMP	0.45	1.6 (1.1-2.3)	5.5	0.019
minPAIN	-0.011	0.99 (0.98-1)	6.7	0.0094
minSYMP	-0.016	0.98 (0.97-0.99)	9.9	0.0017
PREV.OA	0.35	1.4 (1-2)	4.3	0.038
GEN.A	0.28	1.3 (0.94-1.8)	2.6	0.11
GENDER	0.33	1.4 (0.96-2)	3	0.082

	coef	exp(coef)	se(coef)	z	p
BMIBMI_25-29.9	0.4760	1.6096	0.2485	1.915	0.0554
BMIBMI_30+	1.1486	3.1537	0.2397	4.792	1.65e-06

Likelihood ratio test=27.65 on 2 df, p=9.882e-07
n= 1047, number of events= 139

Figure 4-10: Univariate variable importance. The tables differ for BMI as this variable has three levels, and considers each level as a seperate component.

Although several of the variables are determined to be significant by the univariate analysis, there is no consideration for confounding factors and the influence that different variables have on the development of KOA, so further univariate investigations will not be considered.

4.6.2.2.2. Multivariate

The multivariate Cox regression is used to assess how the covariates jointly influence the probability of the subject developing KOA. The significant variables, based on this, that have a strong association with the outcome are given with a $p - value > 0.05$. In this analysis the significant variables are *BMI*, *Gender,* and *previous knee injury.* These findings are shown in Figure 4-11.

Also shown in Figure 4-11 are the significance tests to assess the suitability of the null hypothesis that the beta values, labelled as 'coef', are equal to zero. In this case the three tests, Likelihood ratio, Wald and Logrank Tests all give $p - values > 0.05$, so therefore reject the null hypothesis, as $\beta \neq 0$.

Another measure of model performance is the concordance, or C-statistic. A guess would give a C-statistic of 0.5 and a perfect model would give a C-statistic equal to 1. The model generated using all of the variables for the 7-year OAI cohort gives a C-statistic of 0.719. This is classed as a good model [136].

```
                    coef exp(coef)  se(coef)      z        p
AGE             0.006949  1.006973  0.010215  0.680 0.496328
FAMILY_HYes    -0.375800  0.686740  0.245443 -1.531 0.125742
SYMPTOMSYes     0.141069  1.151504  0.329142  0.429 0.668218
BMIBMI_25-29.9  0.552354  1.737338  0.256454  2.154 0.031255
BMIBMI_30+      1.163688  3.201720  0.250883  4.638 3.51e-06
EV.INJYes       0.851626  2.343454  0.199013  4.279 1.88e-05
HAND.OAYes     -0.283041  0.753489  0.282094 -1.003 0.315688
maxWOMAC        0.007690  1.007719  0.015850  0.485 0.627563
HIST.FALLYes    0.173676  1.189670  0.176503  0.984 0.325122
KNEE.SURGERYYes -0.243247  0.784078  0.251860 -0.966 0.334142
PAIN.MEDSYes    -0.007162  0.992863  0.292325 -0.025 0.980453
B.LINE_SYMPYes  0.209084  1.232549  0.225042  0.929 0.352842
minPAIN         0.006790  1.006814  0.013879  0.489 0.624666
minSYMP        -0.003849  0.996158  0.008412 -0.458 0.647223
PREV.OAYes      0.203353  1.225506  0.328974  0.618 0.536480
GEN.AYes        0.120827  1.128430  0.321472  0.376 0.707025
GENDERFemale    0.749391  2.115712  0.214876  3.488 0.000487

Likelihood ratio test=66.3  on 17 df, p=9.275e-08
n= 1047, number of events= 139

    Concordance= 0.719  (se = 0.024 )
    Likelihood ratio test= 66.3  on 17 df,   p=9e-08
    Wald test            = 64.18  on 17 df,   p=2e-07
    Score (logrank) test = 68.01  on 17 df,   p=5e-08
```

Figure 4-11: The Multivariate Cox regression output and the statistical test results on the model.

The forest plot in Figure 4-14 visually shows the information relating to the hazard ratios in Figure 4-11. It can be seen that variables contribute significantly to the development of clinical KOA. For variables that are not significant, the confidence bars on the hazard ratios cross the '1' line, indicating the variables lack significance to the development of clinical KOA in this given window.

The next step of the analysis was to test the proportional hazards assumptions.

Figure 4-12 shows the outcome of the test assessing if each variable fits the proportional hazards assumption. In the case of this model, the test for each individual variable is not statistically significant so the assumption of proportional hazards for the covariates holds. However, the final item in the list, Global, has a $p - value = 0.024$, a significant result. Despite the significant p-value result, this result can in part be overlooked, as the

significance could be related to the size of the sample, $(n = 1003)$ where large values of n can make p-values less reliable. This is also not a concern that the global option is significant as this is not the final model being used. This model, containing all of the covariates, is a baseline assessment of the features in the data and will be subject to feature selection methods, where if a variable is still showing significance other steps will be taken to allow for this in the calculations.

The survival curves in Figure 4-13 is the unstratified curve for the training and test sets. The test curve shows the predictions anticipate a lower survival than the actual data, but the actual data lies within the confidence interval for the predictions, so therefore is valid. The wider band of the confidence interval is toward the end of the window, from about 5 years, where there are fewer observations within that class.

	chisq	df	p
AGE	0.05530	1	0.814
FAMILY_H	0.65041	1	0.420
SYMPTOMS	0.78929	1	0.374
BMI	0.58364	2	0.747
EV.INJ	3.26881	1	0.071
HAND.OA	0.31757	1	0.573
maxWOMAC	3.03709	1	0.081
HIST.FALL	1.10529	1	0.293
KNEE.SURGERY	2.48440	1	0.115
PAIN.MEDS	0.00207	1	0.964
B.LINE_SYMP	1.66657	1	0.197
minPAIN	0.84218	1	0.359
minSYMP	0.18151	1	0.670
PREV.OA	0.84768	1	0.357
GEN.A	0.05249	1	0.819
GENDER	0.71666	1	0.397
GLOBAL	30.28827	17	0.024

Figure 4-12: The results of the tests for the model assumption of proportional hazards.

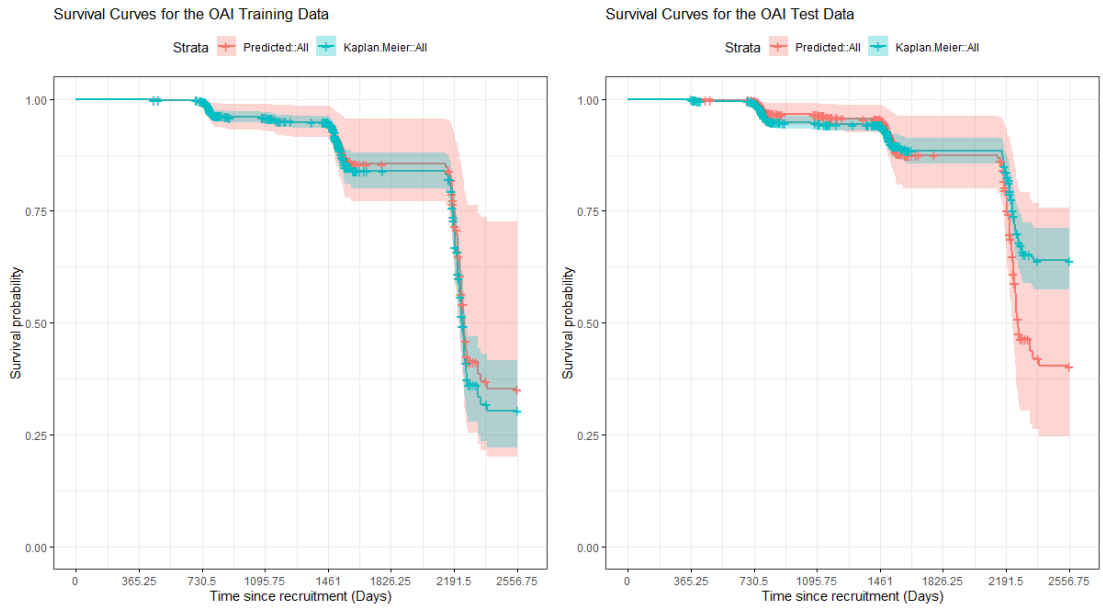


Figure 4-13: The survival curves for the models using both the training and the test data for the whole covariate selection, and compared to the unadjusted curve, for the model considering the 7-year cohort with all variables in the model.

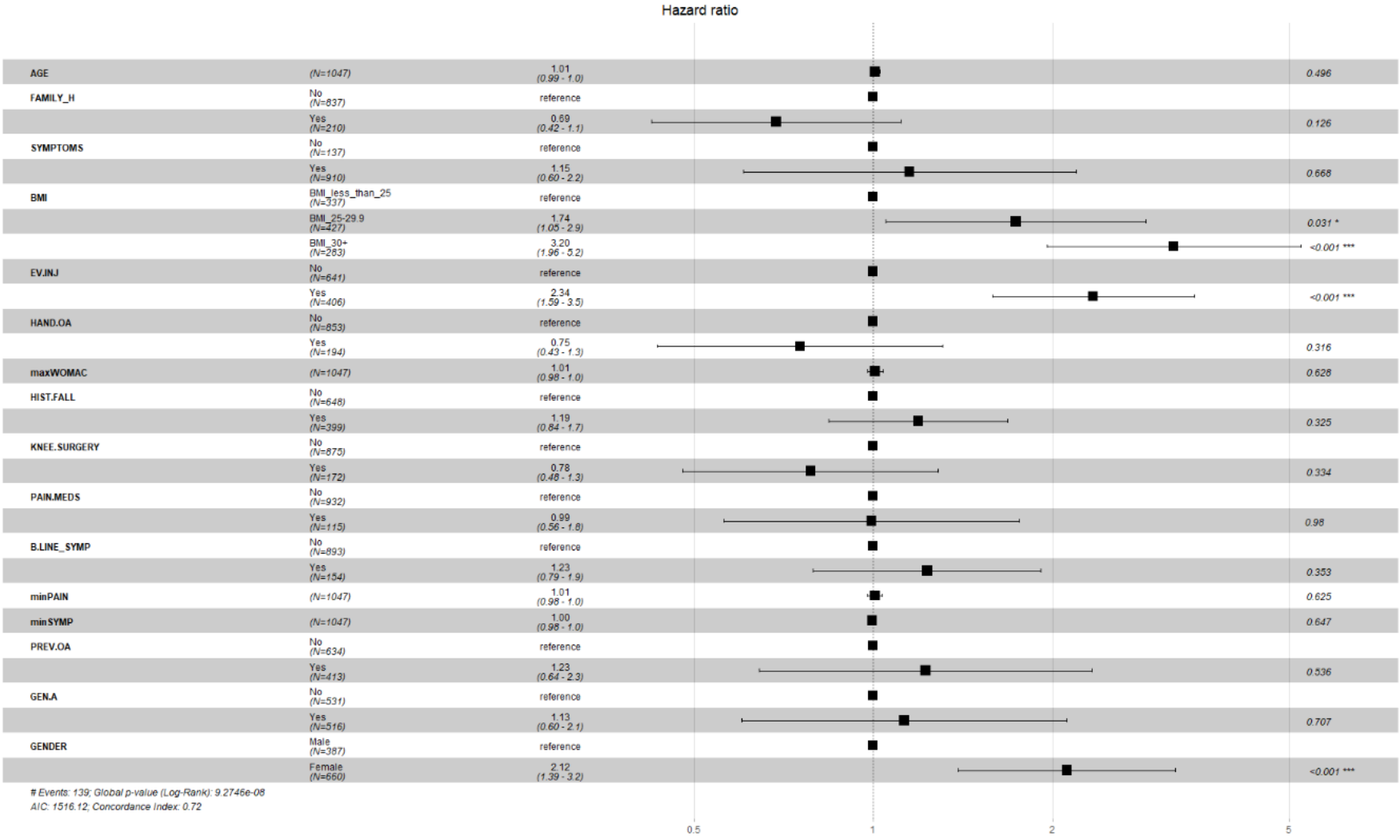


Figure 4-14: Forest plot showing the hazard ratios for the variables included in the full multivariate Cox model.

4.6.2.2.3. Stepwise Multivariate Cox Regression

The two types of stepwise feature selection used have different approaches. The idea behind using both was to establish what variables are selected in each approach, and in the event the variables differ, see if there is an overlap and run a model with the common variables. This would then be compared to the forward and backward models to see if there was any information lost through the removal of variables.

In this analysis, the forward and backward models end with the same variables in the final models, which results in the same beta values in the Cox model. In both cases the proportional hazards results are consistent.

4.6.2.2.3.1. Forward

In using the forward stepwise feature selection, the criteria for measuring the suitability of the variables included in the model is the Akaike Information Criterion (AIC), with the aim of reducing the AIC value, as the lower score indicates a balance between the fit of the data and its ability to not overfit to the data present. The AIC starting value was 1548.42 and including all 16 variables, with the final AIC value of 1499.96 including only 5 variables.

Figure 4-15 depicts the Cox regression output for the forward stepwise model. The variables included were BMI, previous history of knee injury (EV.INJ), gender, family history of knee issues (FAMILY_H) and previous OA diagnosis in other joints (PREV.OA).

Although there are five variables selected by this model only three are significant by the use of $p - values$. The significant variables are BMI, gender and EV.INJ. Even though some variables are not significant the model is the correct model. There is the potential for an element of overfitting due to this despite the stepwise model working as it should, by reducing the AIC value.

Figure 4-15 also shows the statistical tests, Likelihood ratio, Wald and Logrank tests. All three of the tests are significant as their p-values are less than 0.05. This means that the null hypothesis, that the $\beta = 0$, can be rejected as the statistics are similar.

When looking at Figure 4-15, it appears that a family history of KOA lowers the rate of onset when compared to those with no history. This may initially seem counterintuitive but knowing there is a disease in the family that has some modifiable risk factors may cause people to take more care in situations that would increase the risk of developing

KOA. This is seen also in people who know cancer runs in the family, so they take advantage of screening and genetic tests.

```
                   coef exp(coef) se(coef)      z        p
BMIBMI_25-29.9  0.5802    1.7864   0.2518  2.304 0.021221
BMIBMI_30+      1.2176    3.3789   0.2423  5.024 5.06e-07
EV.INJYes       0.8402    2.3169   0.1831  4.589 4.45e-06
GENDERFemale    0.7674    2.1542   0.2032  3.778 0.000158
FAMILY_HYes    -0.4040    0.6676   0.2367 -1.707 0.087914
PREV.OAYes      0.2994    1.3490   0.1716  1.745 0.080979

Likelihood ratio test=60.46  on 6 df, p=3.622e-11
n= 1047, number of events= 139

  Concordance= 0.699  (se = 0.026 )
  Likelihood ratio test= 60.46  on 6 df,   p=4e-11
  Wald test            = 57.91  on 6 df,   p=1e-10
  Score (logrank) test = 60.66  on 6 df,   p=3e-11
```

Figure 4-15: The Multivariate Cox regression output and the statistical test results on the model for the model resulting from the forward stepwise Cox regression model.

The data presented in Figure 4-15 is shown graphically in Figure 4-16, with the use of a forest plot. The variables chosen in this analysis are indicated on the left along with the respective significance for each variable on the right. The bars show the confidence intervals for the hazard ratios for each variable.

Figure 4-16 makes it easier to see the features that contribute to the development of KOA. For example, having a previous knee injury can be linked to the onset of KOA. Being female puts a person at increased risk of developing KOA. Also, any BMI above the NHS 'healthy' limit of 25 is associated with an increased risk of developing KOA, with BMI 30+ showing a greater risk for developing KOA.

Figure 4-17 is the results from the proportional hazards assumption tests. All of the covariates have p-values greater than 0.05, meaning that none breach the proportional hazards assumption. In addition, the global model is not significant, so the proportional hazards assumption holds on the cumulative use of the covariates in the model.

Figure 4-18 shows the Schoenfeld test, which is the graphical test of proportional hazards. The graphs show what has been presented in Figure 4-17. Each covariate supports the assumption of the proportional hazards.

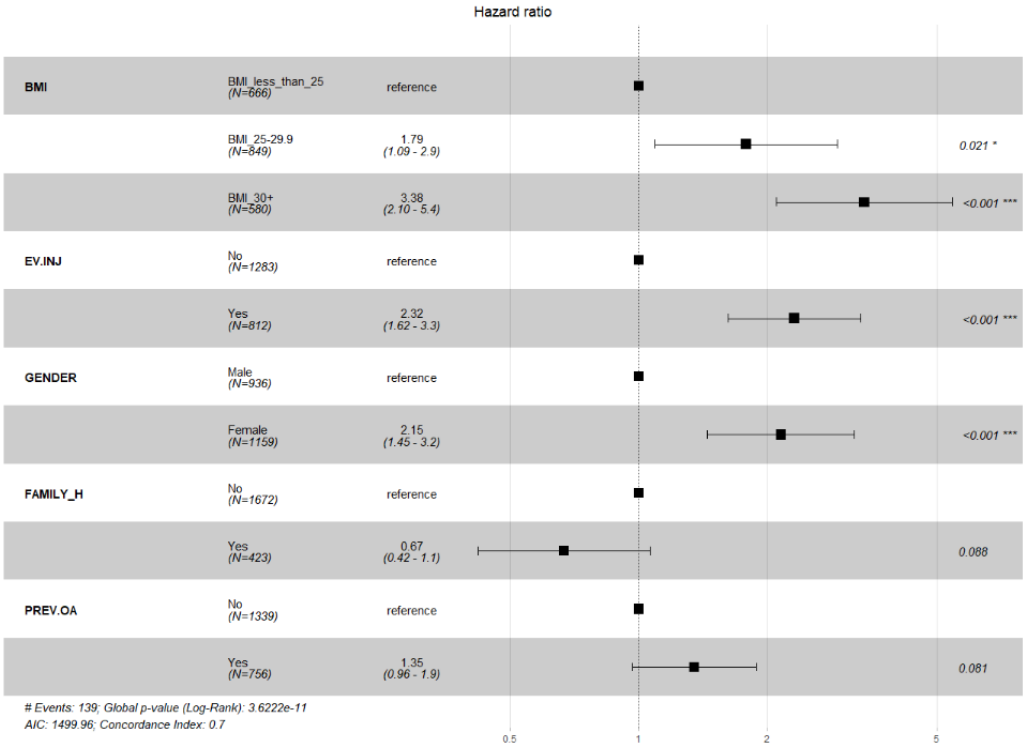


Figure 4-16: Forest plot showing the hazard ratios for the variables included in the forward stepwise multivariate Cox model.

	chisq	df	p
BMI	0.773	2	0.68
EV.INJ	2.548	1	0.11
GENDER	0.686	1	0.41
FAMILY_H	0.858	1	0.35
PREV.OA	1.028	1	0.31
GLOBAL	5.925	6	0.43

Figure 4-17: The results from the test for the proportional hazards assumptions. All covariates have $p-values$ greater than 0.05, so none are significant.

When comparing the survival curves in Figure 4-19, the one on the left shows the training curves and the right, the test curves. In the left diagram, the model predictions are close to actual values throughout the plot. In the right, the predictions for the test set are about 25% lower than the actual test set values. When looking at the curves closer, the gap between actual and prediction is close up to about the 2200-day mark, when the curves begin to diverge. However, even with this difference the two curves still overlap slightly in the confidence intervals.

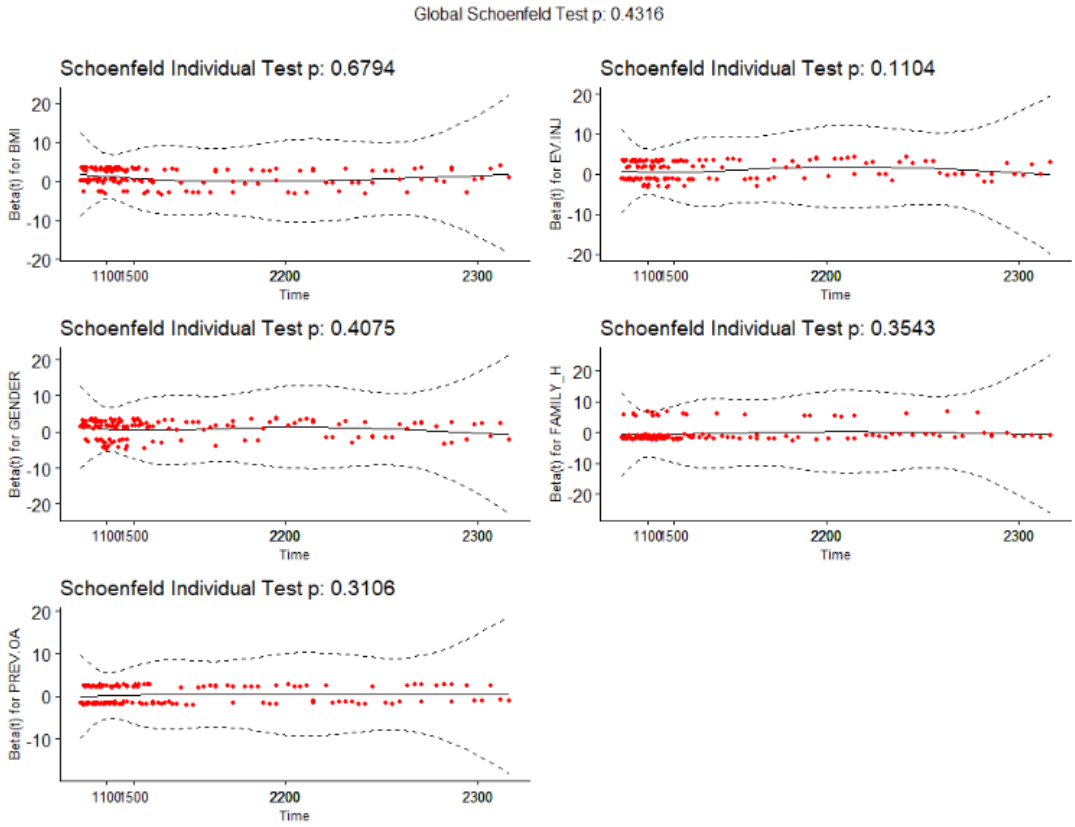


Figure 4-18: The graphs showing the Schoenfeld test on each variable to check the proportional hazards assumption.

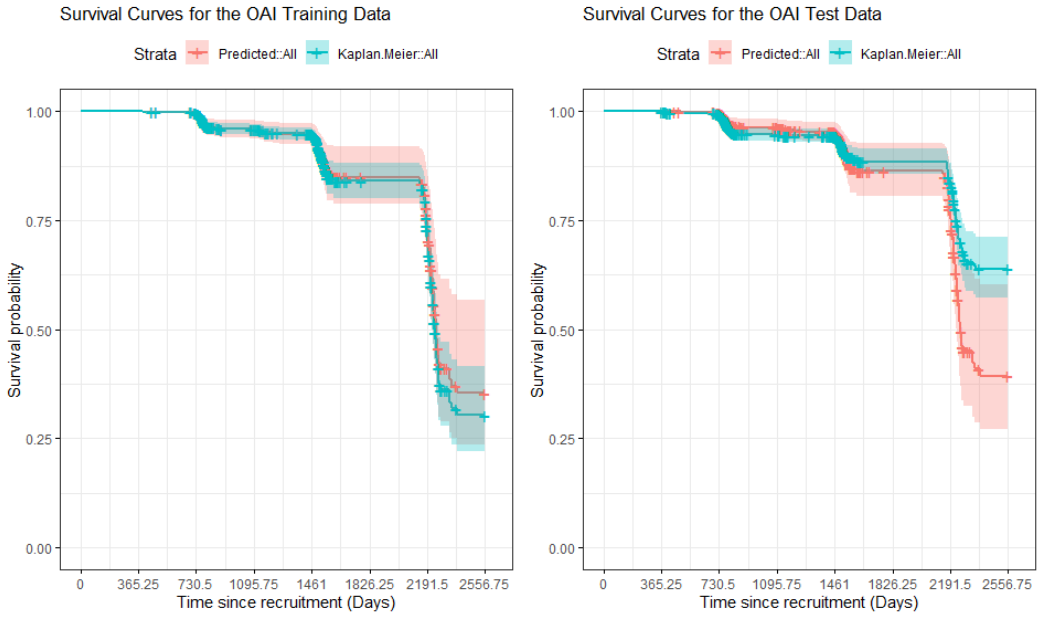


Figure 4-19: The survival curves for the models using both the training and the test data for the covariate subset selected using forward stepwise selection and compared to the unadjusted curve.

4.6.2.2.3.2. Backward

In using the backward stepwise feature selection, the criteria for measuring the suitability of the variables included in the model is the Akaike Information Criterion (AIC), with

the aim of reducing the AIC value, as the lower score indicates a balance between the fit of the data and its ability to not overfit to the data present. The AIC starting value was 1516.12 and initialising with no variables, with the final AIC value of 1499.96 including only five variables. The five variables in the final backward stepwise Cox regression model are the same five that were selected using the forward stepwise Cox regression model.

The variables are shown in Figure 4-20 are BMI, family history of knee issues (FAMILY_H), previous injury to the knee (EV.INJ), previous OA diagnosis in another joint in the body (PREV.OA) and gender.

Although there are five variables selected by this model only three are significant by the use of $p - values$. The significant variables are BMI, gender and EV.INJ, which are the same variables in the forward stepwise model. Even though some variables are not significant the model is the correct model. There is the potential for an element of overfitting due to this despite the stepwise model working as it should, by reducing the AIC value.

Figure 4-20 also shows the statistical tests, Likelihood ratio, Wald and Logrank tests. All three of the tests are significant as their p-values are less than 0.05. This means that the null hypothesis, that the $\beta = 0$, can be rejected as the statistics are similar. The model concordance is 0.699 (0.70 two d.p.) which means that this is classified as a good model. Looking at the concordance scores, there has not been a large drop between the model with all the variables and the model only considering the smaller subset. Therefore, it would be worth the drop in c-statistic for the increase in simplicity brought on from the simpler model only containing 5 variables.

```
                  coef exp(coef) se(coef)      z        p
BMIBMI_25-29.9  0.5802    1.7864   0.2518  2.304 0.021221
BMIBMI_30+      1.2176    3.3789   0.2423  5.024 5.06e-07
FAMILY_HYes    -0.4040    0.6676   0.2367 -1.707 0.087914
EV.INJYes       0.8402    2.3169   0.1831  4.589 4.45e-06
PREV.OAYes      0.2994    1.3490   0.1716  1.745 0.080979
GENDERFemale    0.7674    2.1542   0.2032  3.778 0.000158

Likelihood ratio test=60.46  on 6 df, p=3.622e-11
n= 1047, number of events= 139

    Concordance= 0.699  (se = 0.026 )
    Likelihood ratio test= 60.46  on 6 df,   p=4e-11
    Wald test            = 57.91  on 6 df,   p=1e-10
    Score (logrank) test = 60.66  on 6 df,   p=3e-11
```

Figure 4-20: The Multivariate Cox regression output and the statistical test results on the model for the model resulting from the backward stepwise Cox regression model.

Figure 4-20 shows the hazard ratios calculated from using the Backard stepwise Cox

regression model. The variables here are the same as the ones selected from forward stepwise Cox regression. As the results for the forward and backward stepwise models are the same, the forest plot in Figure 4-16 also applies here.

To ensure the modelling of the variables is appropriate for the assumptions made about proportional hazards, a test is performed. Based on the results shown in Figure 4-21 all of the covariates, along with the model as a whole, follow the proportional hazards assumption.

	chisq	df	p
BMI	0.773	2	0.68
FAMILY_H	0.858	1	0.35
EV.INJ	2.548	1	0.11
PREV.OA	1.028	1	0.31
GENDER	0.686	1	0.41
GLOBAL	5.925	6	0.43

Figure 4-21: The results from the test for the proportional hazards assumptions on the covariates from backward stepwise Cox regression. All covariates have p-values greater than 0.05, so none are significant.

4.6.2.2.4. Stepwise with Stratification

As the forward and backward models both give the same variables in for the Cox regression, the model variables are consistent. The variables are BMI, family history of knee issues (FAMILY_H), previous injury to the knee (EV.INJ), previous OA diagnosis in another joint in the body (PREV.OA) and gender.

The next step in the analysis is to see if there are groups within the cohort. For example, to determine if there is a high and low risk group, and to establish what the criteria is for inclusion in each group.

To stratify the group into cohorts the first step is to establish a cut point that gives the biggest separation in the subjects. Figure 4-22 shows the histogram along with the cut point that is used to define the two strata that will produce a high and low risk cohort. The process of the stratification is discussed earlier in section 4.4.7.

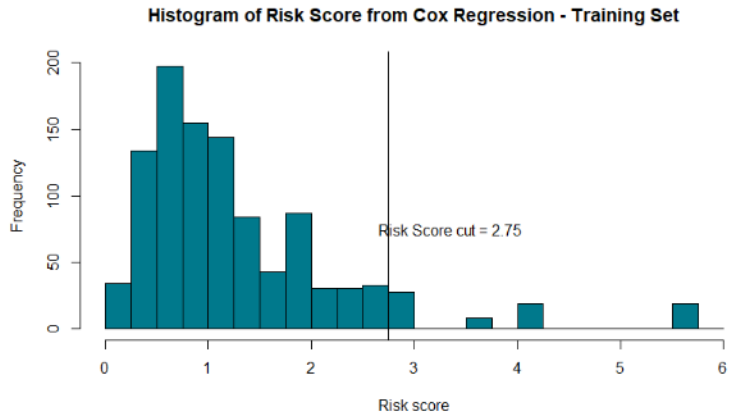


Figure 4-22: Histogram of the risk score for the training data when establishing where to add the cut point. In this case, the cut point falls

where the risk score is 2.75.

Figure 4-23 shows the survival curves for the training data after it has been split into the risk stratification cohorts. The left curve shows the actual data and how it is divided. Note, that in cohort 2 the last recorded event occurs within the 7-year span at day 2211. The curve on the right shows the predictions once they have been split into the risk stratification cohorts. As there is no overlap of the confidence intervals, the two populations are different from each other. Therefore, the idea of two risk groups is valid for this dataset.

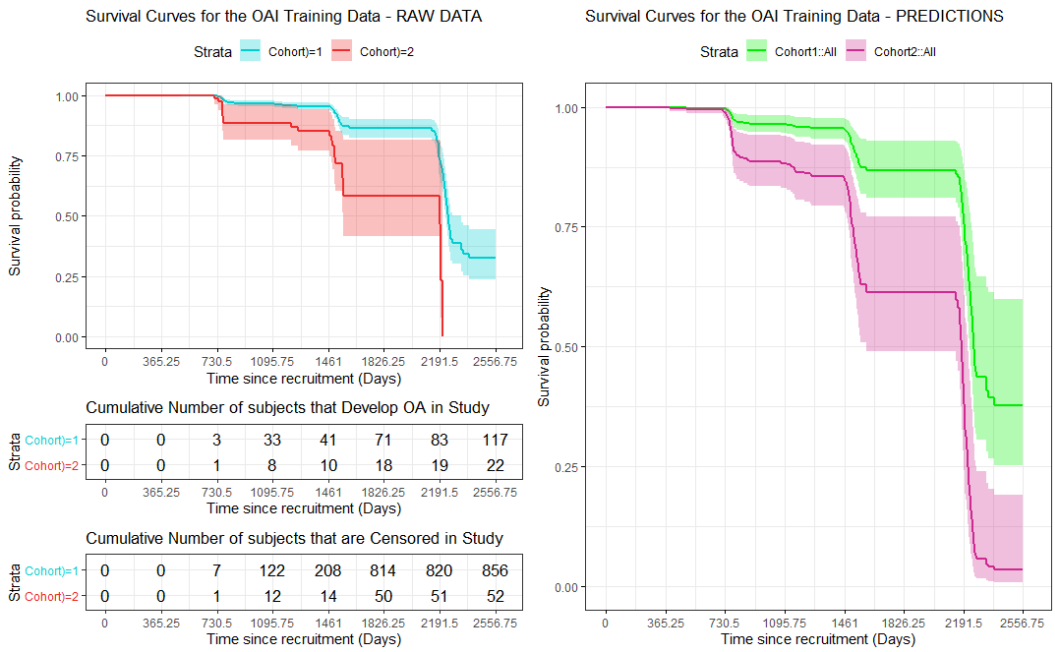


Figure 4-23: Stratification curves on the training data showing the high and low risk cohorts. Note that in cohort 2 the last event recorded in cohort 2 within the 7-year span is at day 2211.

The histogram in Figure 4-24 is showing the risk scores from the test set and the predetermined cut point and the way this divides the data. Figure 4-25 shows the stratification curves on the test data for both the raw data and the predictions produced on the test set. The stratification shows an overlap in the survival curves on the test data. The large confidence interval on Cohort 2 could be due to the smaller cohort size when compared to that of Cohort 1. The predictions produce curves that do not overlap.

Figure 4-24: Histogram of the risk score for the test data showing where the predetermined cut of the risk score falls.

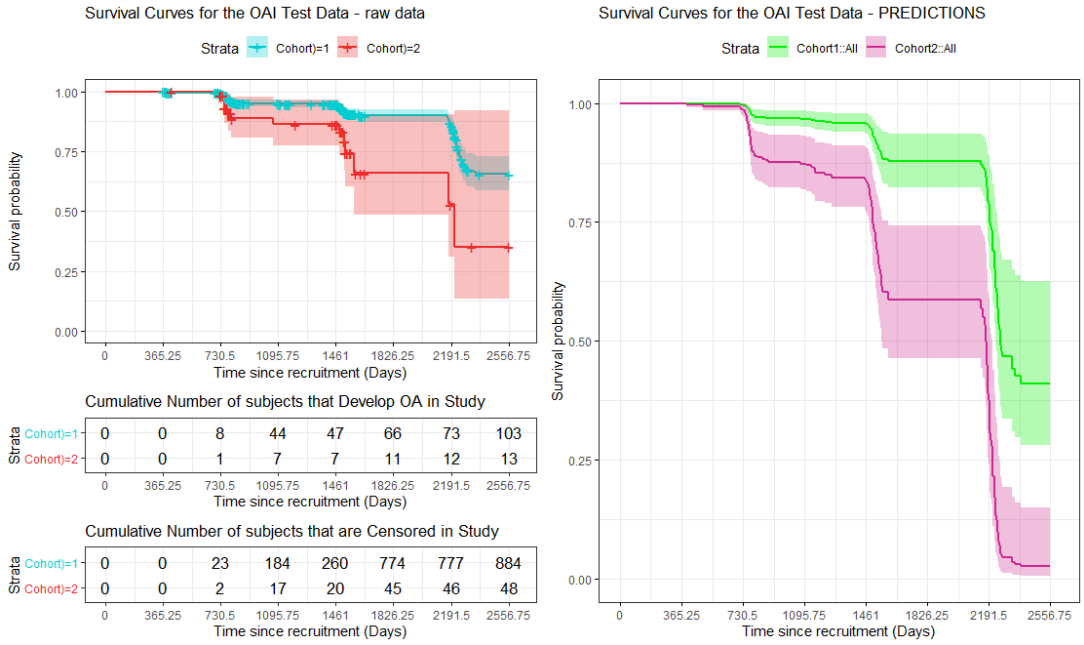


Figure 4-25: Stratification curves on the test data showing the high and low risk cohorts using the predetermined cut points calculated on the training data.

For the model to have clinical value the findings of the two risk cohorts need to be translated into human terms. For example how the features influence that individual in relation to which risk group they will belong. The proportions of each cohort by covariate are shown in Figure 4-26.

The proportion plots are useful as they can be used easier to profile the groups in each cohort. For example, in Cohort 2 the majority of the subjects are female with a BMI over 25, and have likely had previous knee injuries. The majority of the people in cohort 2 have no family history of knee problems, which could mean that those who are aware of the issues their families had already made changes to their behaviours to help prevent them from developing KOA.

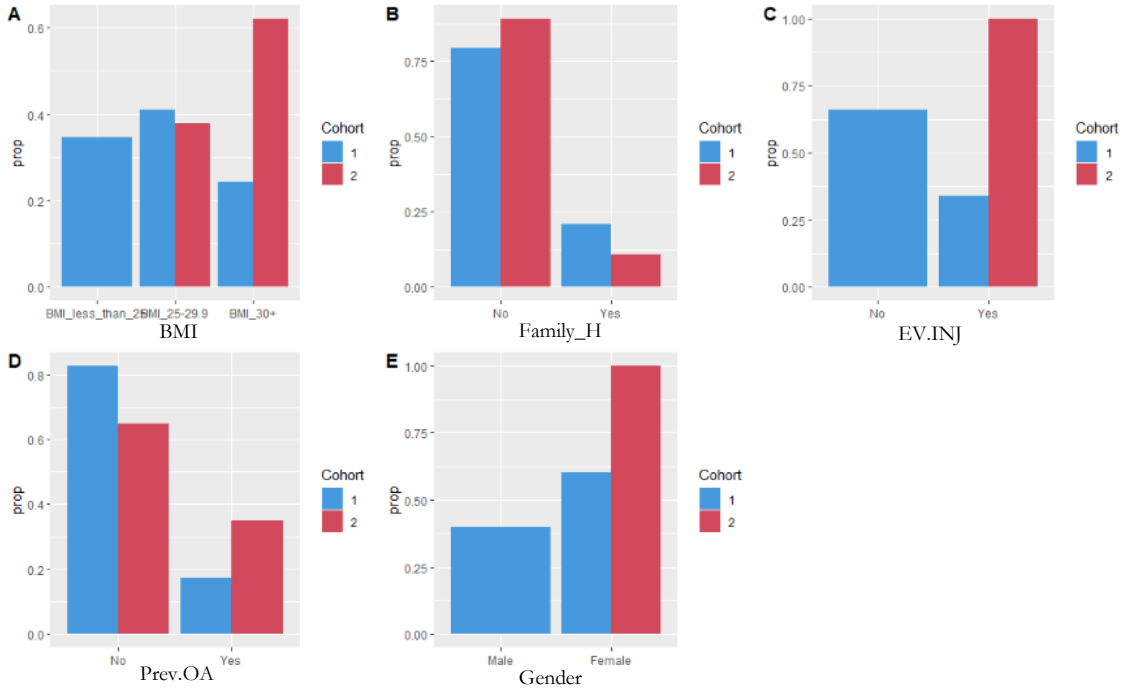


Figure 4-26: The cohort profiles per variable for the different strata. The blue bars show cohort 1 and red show cohort 2. This representation of the profiles is the proportion of the group in each data category per cohort for the training set.

The cohort profile for the test data is in Figure 4-27. The same patterns that are evident in the training set are also present in the test set.

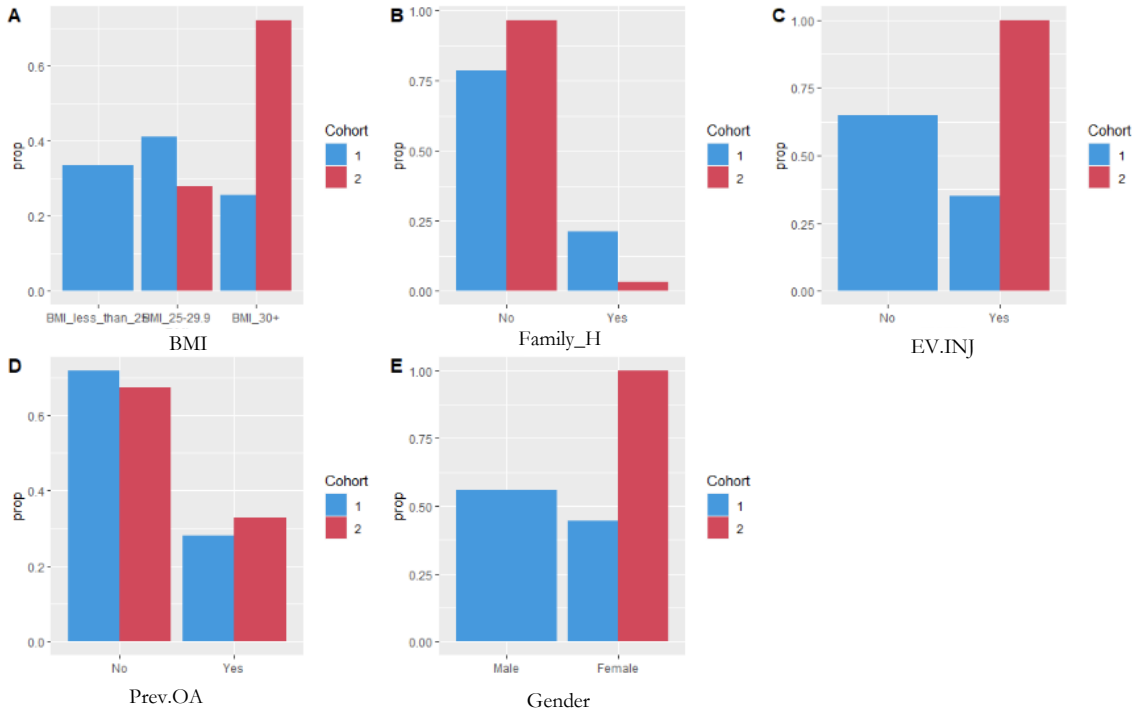


Figure 4-27: The cohort profiles per variable for the different strata. The blue bars show cohort 1 and red show cohort 2. This representation of the profiles is the proportion of the group in each data category per cohort for the test set.

4.6.2. Results from Five year cohort

The stages for the five-year cohort study follow the same layout as that used in the seven-year study.

The five-year study is the approach that has the most potential use as when comparing the survival curves as the difference between the training and test sets are not too different and fall within 10% of each other, with the confidence intervals crossing. This is the model that has potential clinical use.

4.6.2.1. Kaplan Meier Curves

The five-year cohort theoretically has more clinical significance than the seven-year study. This is because the curves from the seven-year analysis, showing training and test sets are quite close up to the five-year mark, when they begin to differ drastically, and the KM curves for the 5-year cohort show a more acceptable distance of separation at about 10%.

Figure 4-28 is the observational KM curve for the five-year cohort considering all of the data. The data appears to show 'steps' where subjects are likely to develop KOA, however as this data was collected as part of a study that had given windows for follow up, these may just be an artefact of the data. This curve, similar to those shown in Figure 4-29 show the overall survival to be around 85% after a five-year follow up.

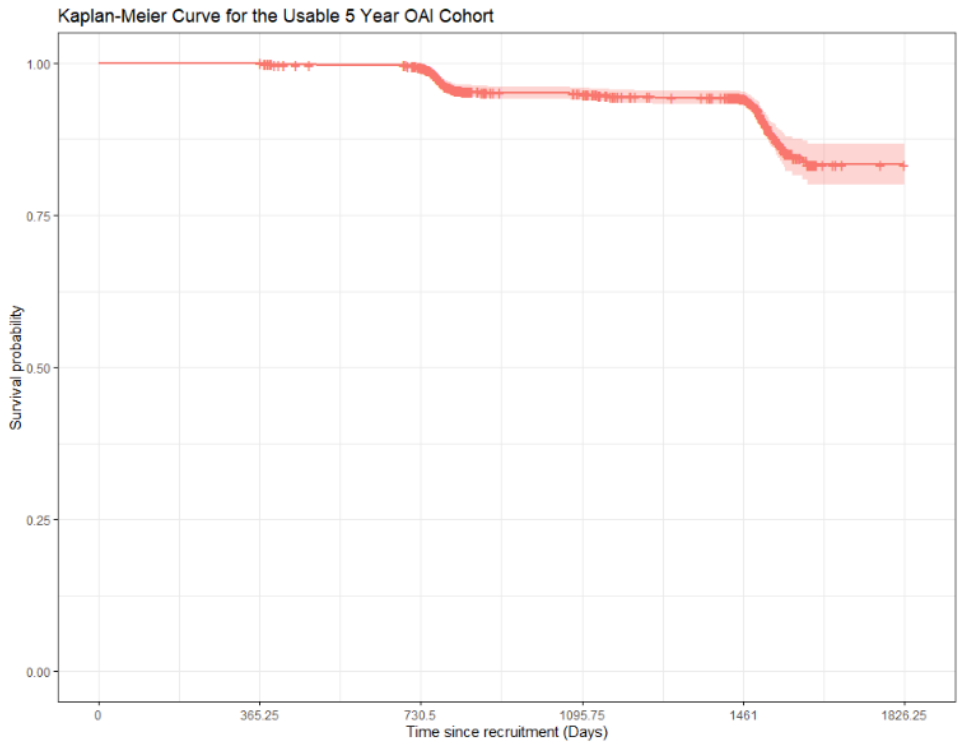


Figure 4-28: An observational Kaplan-Meier curve for the data of the 5-year cohort.

Figure 4-29 shows the KM curves for the whole cohort, training, and test sets for the

five-year window. The red line is the whole cohort KM curve, the green shows the training curve, and the blue shows the test curve. There is a difference between the training and test sets with the survival probability differing between the two groups by about 10%. However, the confidence intervals cross over each other so the divide between the groups is still acceptable.

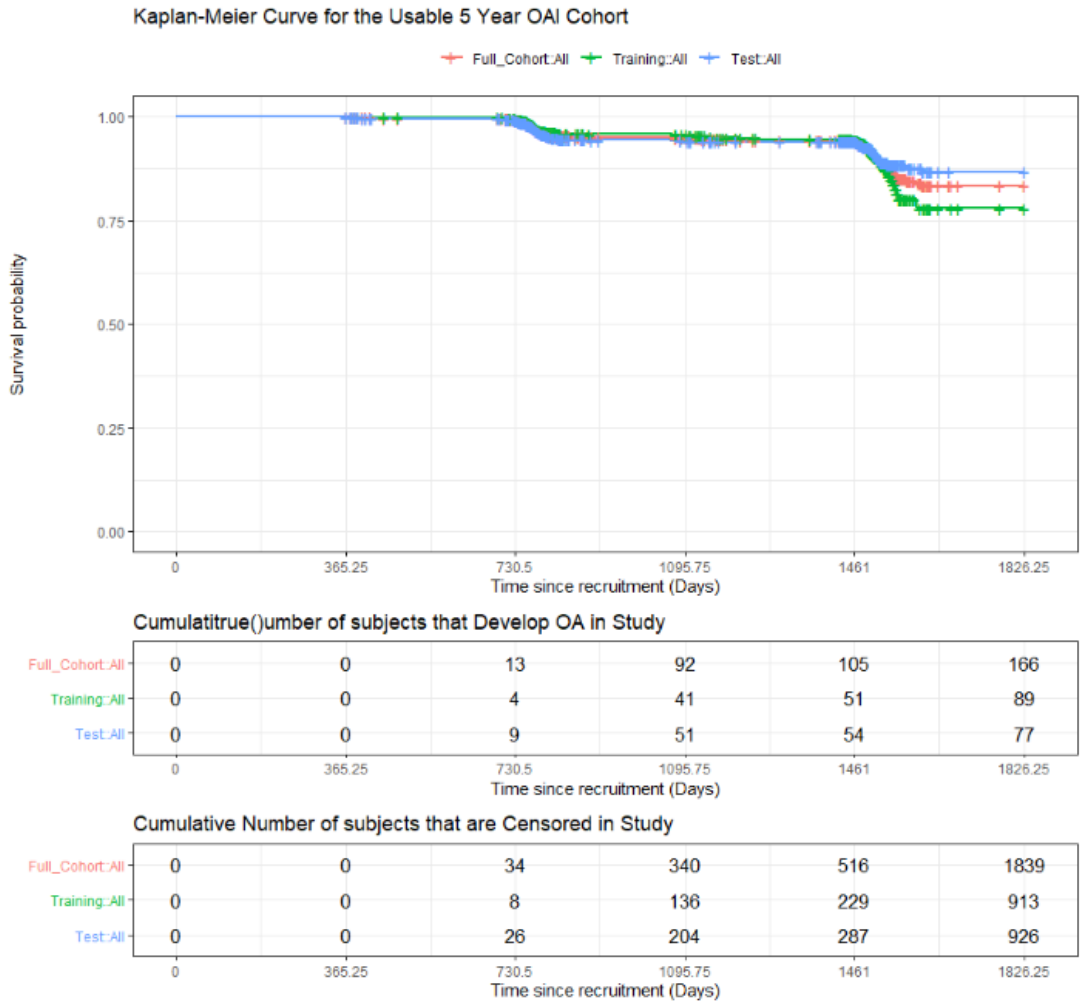


Figure 4-29: Observational KM curve stratified by sample. The red depicts the total sample of the 2005 subjects in the study. The green shows the training sample, and the blue shows the test sample. The tables below illustrate the way in which the data is split between the samples.

4.6.2.2. Cox Regression

4.6.2.2.1. Univariate

The variables are all modelled individually to determine which, if any, are alone significant in development of KOA for the five-year window. The results are shown in Figure 4-30. Based on this, the following variables are significant to development; Ever have a knee injury (EV.INJ), maxWOMAC, history of falling (HIST.FALL), the subjects pain score on the KOOS scoring system (minPAIN), the subjects symptom score on the KOOS scoring system (minSYMP), the subjects gender and their BMI.

However despite there being several variables that are significant in the univariate analysis, there is no consideration for confounding factors and the influence that different variables have on the development of KOA, so further univariate investigations will not be considered.

	beta	HR (95% CI for HR)	wald.test	p.value
AGE	-0.003	1 (0.97-1)	0.06	0.8
FAMILY_H	-0.39	0.68 (0.38-1.2)	1.7	0.2
SYMPTOMS	0.53	1.7 (0.82-3.5)	2	0.15
EV.INJ	0.56	1.7 (1.2-2.6)	6.9	0.0086
HAND.OA	-0.07	0.93 (0.53-1.6)	0.06	0.8
maxWOMAC	0.023	1 (1-1)	14	0.00019
HIST.FALL	0.47	1.6 (1.1-2.4)	5	0.026
KNEE.SURGERY	0.079	1.1 (0.62-1.9)	0.08	0.78
PAIN.MEDS	0.15	1.2 (0.62-2.2)	0.21	0.65
B.LINE_SYMP	0.5	1.6 (0.99-2.7)	3.7	0.053
minPAIN	-0.015	0.98 (0.97-1)	7.7	0.0055
minSYMP	-0.018	0.98 (0.97-0.99)	8.6	0.0034
PREV.OA	0.25	1.3 (0.84-2)	1.4	0.24
GEN.A	0.25	1.3 (0.85-1.9)	1.4	0.24
GENDER	0.5	1.6 (1-2.7)	4.3	0.039

	coef	exp(coef)	se(coef)	z	p
BMIBMI_25-29.9	0.7531	2.1236	0.3386	2.224	0.0261
BMIBMI_30+	1.5825	4.8670	0.3250	4.869	1.12e-06

Likelihood ratio test=31.04 on 2 df, p=1.822e-07
n= 1002, number of events= 89

Figure 4-30: Univariate variable importance. The tables differ for BMI as this variable has three levels, and considers each level as a seperate component.

4.6.2.2.2. Multivariate

The multivariate Cox regression is used to assess how the covariates jointly influences the probability of the subject developing KOA. The significant variables, based on this, that have a strong association with the outcome are given with a $p - value > 0.05$. In this analysis the significant variables are *BMI*, *Gender,* and *previous knee injury.* These variables are the same as those found to be of significance in the 7-year study. These findings are shown in Figure 4-31.

Also shown in Figure 4-31 are the significance tests to assess the suitability of the null hypothesis that the beta values, labelled as 'coef', are equal to zero. In this case the three tests, Likelihood ratio, Wald and Logrank Tests all give $p - values > 0.05$, so therefore reject the null hypothesis as $\beta \neq 0$.

Another measure of model performance is the concordance, or C-statistic. A guess would give a C-statistic of 0.5 and a perfect model would give a C-statistic equal to 1. The model generated using all of the variables for the 5-year OAI cohort gives a C-statistic of 0.754, an increase on this measure for the 7-year study. This is classed as a good model [136].

```
                     coef exp(coef)  se(coef)      z        p
AGE              0.001422  1.001423  0.013012  0.109  0.91296
FAMILY_HYes     -0.495713  0.609137  0.306202 -1.619  0.10547
SYMPTOMSYes      0.264408  1.302659  0.396589  0.667  0.50496
BMIBMI_25-29.9   0.771648  2.163328  0.344733  2.238  0.02520
BMIBMI_30+       1.526708  4.602997  0.336368  4.539 5.66e-06
EV.INJYes        0.603000  1.827593  0.230980  2.611  0.00904
HAND.OAYes      -0.252123  0.777149  0.339247 -0.743  0.45737
maxWOMAC         0.028357  1.028763  0.018252  1.554  0.12028
HIST.FALLYes     0.336964  1.400688  0.216829  1.554  0.12017
KNEE.SURGERYYes -0.116629  0.889915  0.320281 -0.364  0.71575
PAIN.MEDSYes     0.056276  1.057889  0.329435  0.171  0.86436
B.LINE_SYMPYes   0.230422  1.259131  0.300437  0.767  0.44311
minPAIN          0.014329  1.014432  0.015938  0.899  0.36861
minSYMP          0.004246  1.004255  0.011134  0.381  0.70297
PREV.OAYes       0.034526  1.035129  0.395283  0.087  0.93040
GEN.AYes         0.200373  1.221858  0.374760  0.535  0.59288
GENDERFemale     0.805672  2.238200  0.259128  3.109  0.00188

Likelihood ratio test=61.75  on 17 df, p=5.391e-07
n= 1002, number of events= 89

 Concordance= 0.754  (se = 0.023 )
 Likelihood ratio test= 61.75  on 17 df,   p=5e-07
 Wald test            = 61.08  on 17 df,   p=7e-07
 Score (logrank) test = 65.99  on 17 df,   p=1e-07
```

Figure 4-31: The Multivariate Cox regression output and the statistical test results on the model.

Figure 4-32 shows the information found in Figure 4-31 in a graphical way. The forest plot clearly shows which variables are significant to the development of clinical KOA. For variables that are not significant, the confidence bars on the hazard ratios cross the '1' line, indicating the variables lack of significance to the development of clinical KOA.

The proportional hazards assumption needs to be tested in order to use the standard survival analysis approaches. The results of the proportional hazards tests are in Figure 4-33. In order to say that the assumptions are valid all of the variables need to have a $p - value$ greater than 0.05. In this case, shown in Figure 4-33, minPAIN has a p-value of 0.048. In this situation, this is okay for two reasons. The first being the sample size may slightly skew the effectiveness of the p-value and the second being that this is the model that uses all variables, before any feature selection, which follows in the section '4.6.2.2.3'.

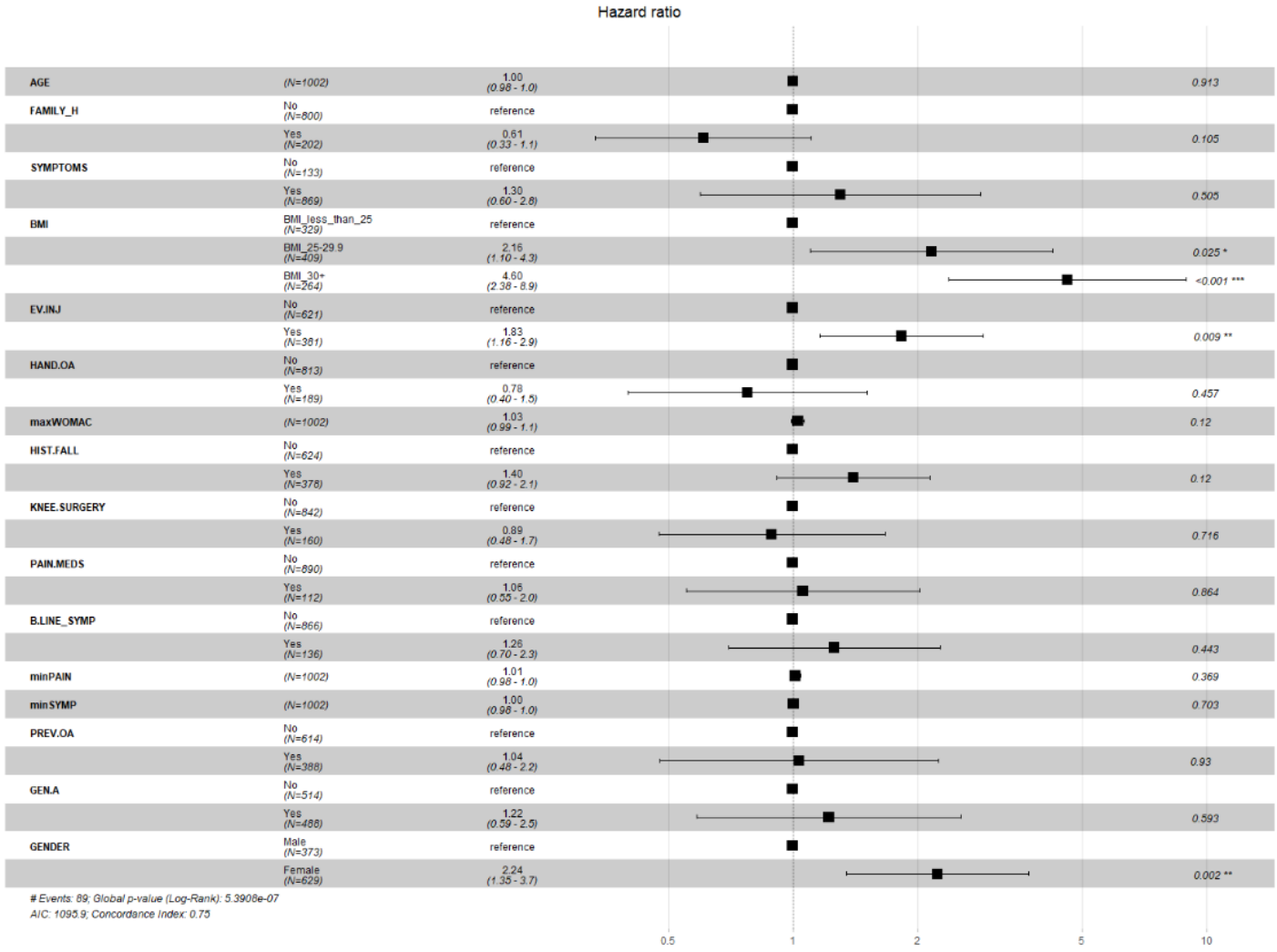


Figure 4-32: Forest plot showing the hazard ratios for the variables included in the full multivariate Cox model.

	chisq	df	p
AGE	0.0051	1	0.943
FAMILY_H	0.3979	1	0.528
SYMPTOMS	0.5103	1	0.475
BMI	5.2679	2	0.072
EV.INJ	0.0766	1	0.782
HAND.OA	0.1277	1	0.721
maxWOMAC	3.3070	1	0.069
HIST.FALL	0.2524	1	0.615
KNEE.SURGERY	0.3430	1	0.558
PAIN.MEDS	0.0862	1	0.769
B.LINE_SYMP	0.0971	1	0.755
minPAIN	3.9235	1	0.048
minSYMP	3.2657	1	0.071
PREV.OA	0.0018	1	0.966
GEN.A	0.7056	1	0.401
GENDER	0.2098	1	0.647
GLOBAL	17.2796	17	0.436

Figure 4-33: The results of the tests for the model assumption of proportional hazards.

The survival curves, shown in Figure 4-34, show the actual and predicted curves for both the training and test sets in the five-year cohort. The confidence intervals on the test graph for the test predictions and the actual test data have an overlap of the confidence intervals, so therefore is a valid result. Again, as for the 7-year cohort, where there are fewer observations within that group, the time between four and five years, the confidence intervals are wider.

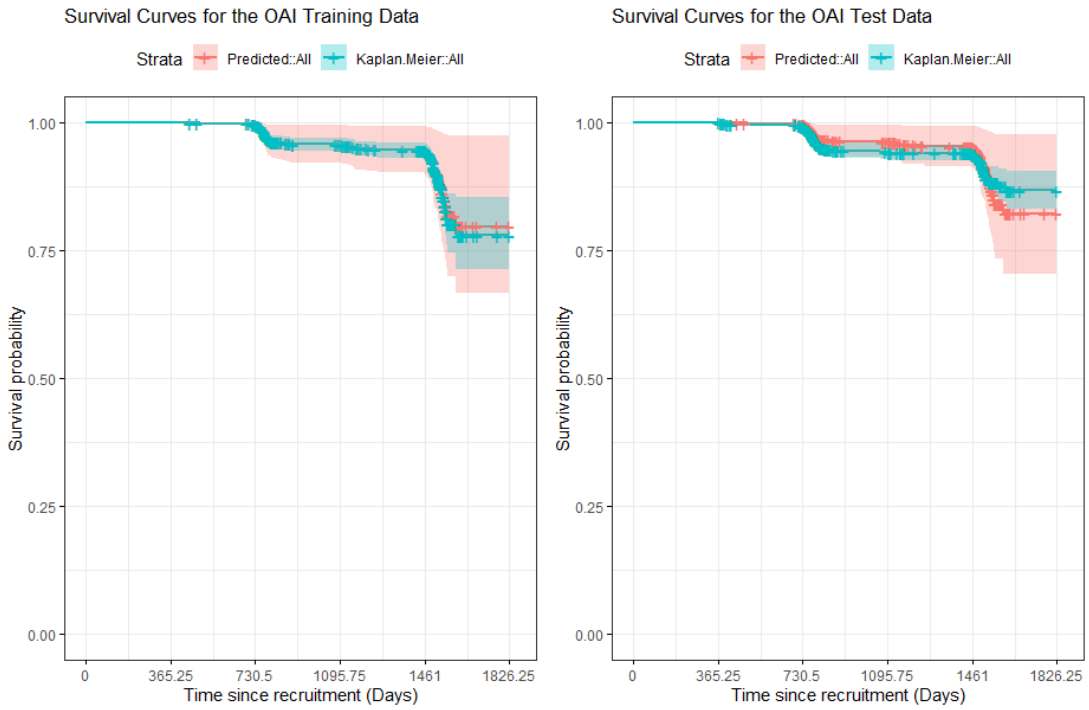


Figure 4-34: The survival curves for the models using both the training and the test data for the whole covariate selection, and compared to the unadjusted curve.

4.6.2.2.3. Stepwise

In the same approach used for the seven-year analysis, two types of stepwise feature

selection have been used. Similar to the seven-year, the forward and backward models end with the same variables in the final models, which results in the same beta values in the Cox model. In both cases the proportional hazards results are consistent.

4.6.2.2.3.1. Forward

The AIC starting value was 1123.65 and including all 16 variables, with the final AIC value of 1079.13 including 6 variables.

Figure 4-35 depicts the Cox regression output for the forward stepwise model. The variables included were BMI, previous history of knee injury (EV.INJ), gender, family history of knee issues (FAMILY_H), a history of falling (HIST.FALL) and the maximum WOMAC score recorded (maxWOMAC).

Although there are six variables selected by this model only four are significant by the use of $p - values$. The significant variables are BMI, gender, EV.INJ and maxWOMAC. Even though some variables chosen for the model are not significant, the AIC is at its lowest value as a result of including non-significant features, family history of knee issues and history of falling. There is the potential for an element of overfitting due to this despite the stepwise model working as it should, by reducing the AIC value.

Figure 4-35 also shows the statistical tests, Likelihood ratio, Wald and Logrank tests. All three of the tests are significant as their p-values are less than 0.05. This means that the null hypothesis, that the $\beta = 0$, can be rejected as the statistics are similar. The concordance statistic is 0.749 (3 d.p.). Looking at the C-statistic for the feature selected model and the model containing all of the variables there is no difference when considering to two decimal places. This means that there has not been a large loss of information by reducing the model size.

```
                    coef exp(coef) se(coef)      z        p
BMIBMI_25-29.9  0.77133   2.16264  0.34193  2.256   0.0241
BMIBMI_30+      1.55718   4.74540  0.33227  4.687 2.78e-06
EV.INJYes       0.54997   1.73320  0.21760  2.527   0.0115
GENDERFemale    0.80547   2.23774  0.24870  3.239   0.0012
FAMILY_HYes    -0.51472   0.59767  0.30238 -1.702   0.0887
HIST.FALLYes    0.35546   1.42684  0.21404  1.661   0.0968
maxWOMAC        0.01463   1.01474  0.00620  2.360   0.0183

Concordance= 0.749  (se = 0.025 )
Likelihood ratio test= 58.52  on 7 df,   p=3e-10
Wald test            = 56.88  on 7 df,   p=6e-10
Score (logrank) test = 61.35  on 7 df,   p=8e-11
```

Figure 4-35: The Multivariate Cox regression output and the statistical test results on the model for the model resulting from the forward stepwise Cox regression model.

The data presented in Figure 4-35 is shown graphically in Figure 4-36. The variables chosen in this analysis are indicated on the left along with the respective significance for

each variable on the right. The bars show the confidence intervals for the hazard ratios for each variable.

Figure 4-36 makes it easier to see the features that contribute to the development of KOA. The same features in the five-year analysis are those that contributed to the development of KOA.

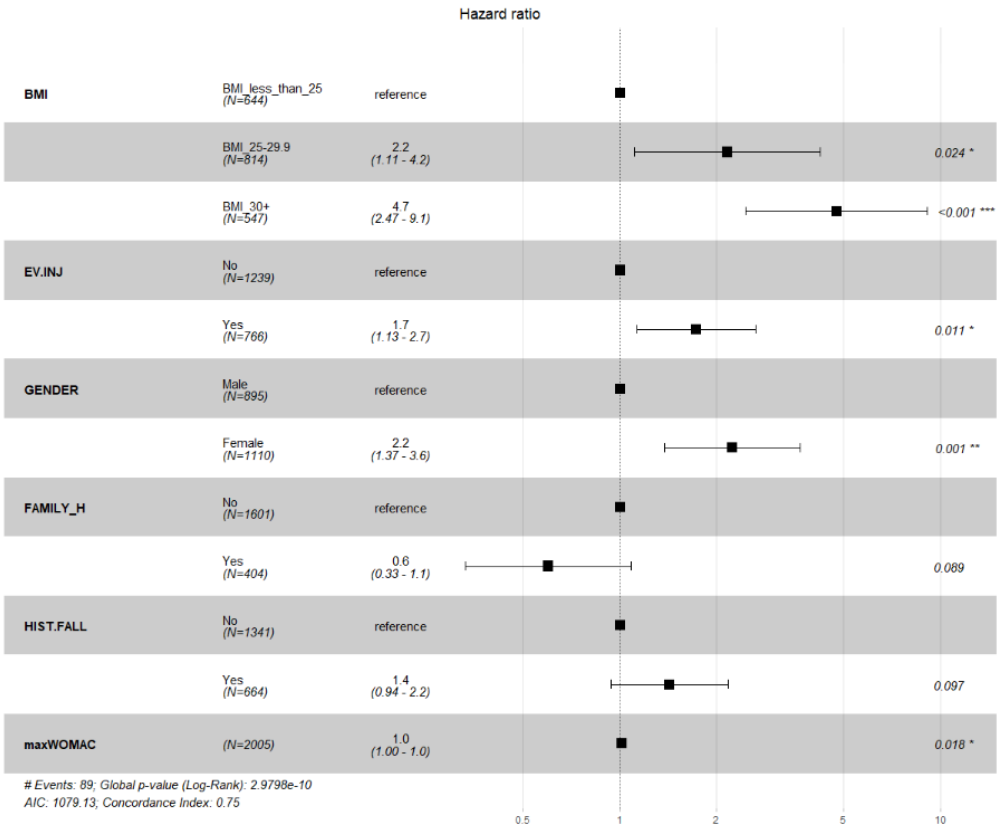


Figure 4-36: Forest plot showing the hazard ratios for the variables included in the forward stepwise multivariate Cox model.

Figure 4-37 is the results from the proportional hazards assumption tests. All of the covariates have p-values greater than 0.05, meaning that none breach the proportional hazards assumption. In addition, the global model is not significant, so the proportional hazards assumption holds on the cumulative use of the covariates in the model. Figure 4-38 show visually the Schoenfeld test for the proportional hazards assumption.

	chisq	df	p
BMI	5.3627	2	0.068
EV.INJ	0.0537	1	0.817
GENDER	0.1446	1	0.704
FAMILY_H	0.3862	1	0.534
HIST.FALL	0.1848	1	0.667
maxWOMAC	3.0387	1	0.081
GLOBAL	9.7857	7	0.201

Figure 4-37: The results from the test for the proportional hazards assumptions. All covariates have p-values greater than 0.05, so none are significant.

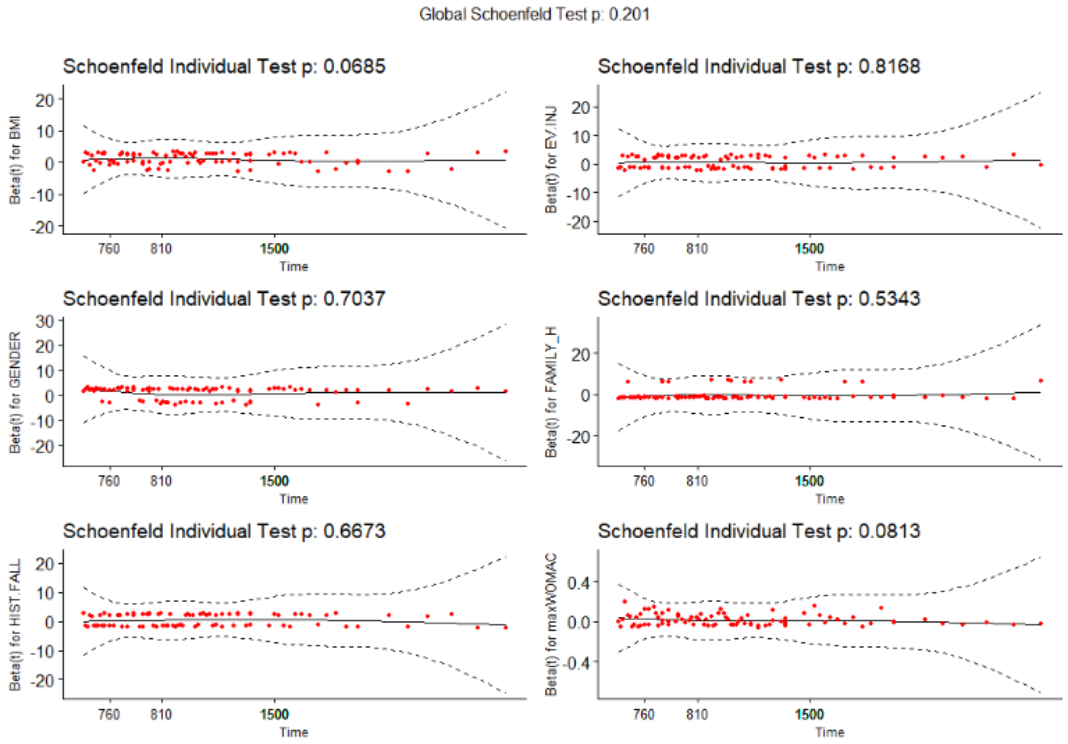


Figure 4-38: The graphs showing the Schoenfeld test on each variable to check the proportional hazards assumption.

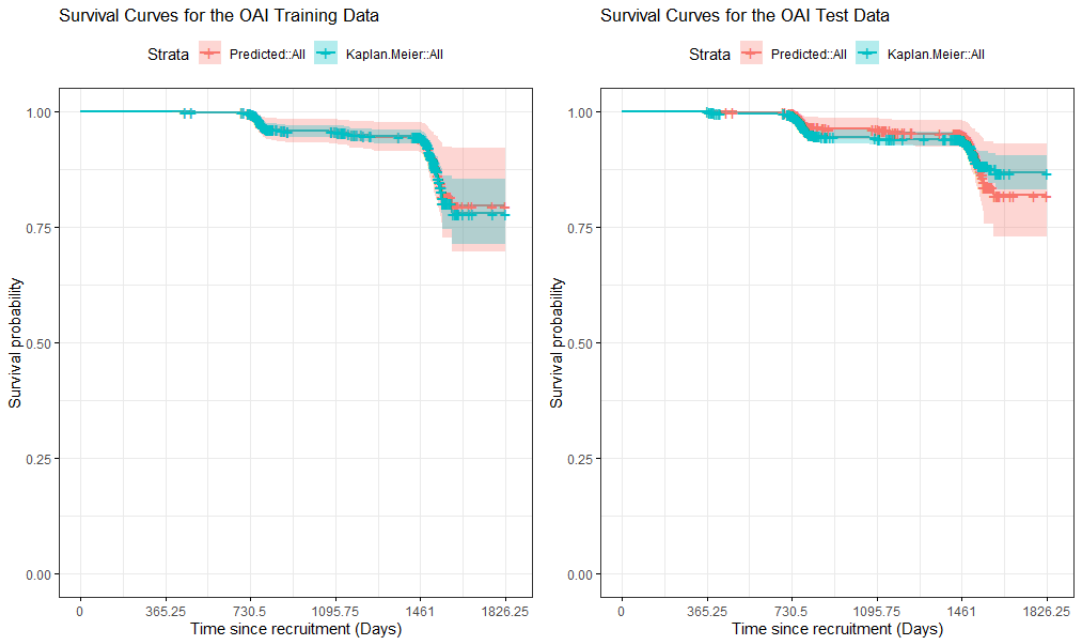


Figure 4-39: The survival curves for the models using both the training and the test data for the covariate subset selected using forward stepwise selection and compared to the unadjusted curve.

When comparing the survival curves in Figure 4-39, the one on the left shows the training curves and the right, the test curves. In the left diagram, the model predictions are close to actual values throughout the plot. In the right, the predictions for the test set are less than 10% lower than the actual test set values. When looking at the curves closer, the gap between actual and prediction is close throughout the time that is being analysed. Even

with this difference, the two curves still overlap in the confidence intervals.

4.6.2.2.3.2. Backward

The AIC starting value was 1095.9 and initialising with all variables, with the final AIC value of 1079.13 including only six variables. The six variables in the final backward stepwise Cox regression model are the same five that were selected using the forward stepwise Cox regression model.

Although there are six variables selected by this model, shown in Figure 4-40, with only four are significant by the use of $p - values$. The significant variables are BMI, gender, maxWOMAC and EV.INJ, which are the same variables in the forward stepwise model. Even though some variables are not significant the model is the correct model. There is the potential for an element of overfitting due to this despite the stepwise model working as it should, by reducing the AIC value.

Figure 4-40 also shows the statistical tests, Likelihood ratio, Wald and Logrank tests. All three of the tests are significant as their p-values are less than 0.05. This means that the null hypothesis, that the $\beta = 0$, can be rejected as the statistics are similar. The model concordance is 0.749 (0.75 two d.p.) which means that this is classified as a good model. Looking at the concordance scores to two decimal places, there has not been any drop between the model with all the variables and the model only considering the smaller subset. Therefore, for the increase in simplicity and retention of information, it makes more sense to use the simpler model only containing six variables.

```
                  coef exp(coef) se(coef)      z        p
BMIBMI_25-29.9 0.77133   2.16264  0.34193  2.256   0.0241
BMIBMI_30+     1.55718   4.74540  0.33227  4.687 2.78e-06
FAMILY_HYes   -0.51472   0.59767  0.30238 -1.702   0.0887
EV.INJYes      0.54997   1.73320  0.21760  2.527   0.0115
maxWOMAC       0.01463   1.01474  0.00620  2.360   0.0183
GENDERFemale   0.80547   2.23774  0.24870  3.239   0.0012
HIST.FALLYes   0.35546   1.42684  0.21404  1.661   0.0968

Likelihood ratio test=58.52  on 7 df, p=2.98e-10
n= 1002, number of events= 89

   Concordance= 0.749  (se = 0.025 )
   Likelihood ratio test= 58.52  on 7 df,   p=3e-10
   Wald test            = 56.88  on 7 df,   p=6e-10
   Score (logrank) test = 61.35  on 7 df,   p=8e-11
```

Figure 4-40: The Multivariate Cox regression output and the statistical test results on the model for the model resulting from the backward stepwise Cox regression model.

Figure 4-41 shows the hazard ratios calculated from using the Backard stepwise Cox regression model. The variables here are the same as the ones selected from forward stepwise Cox regression.

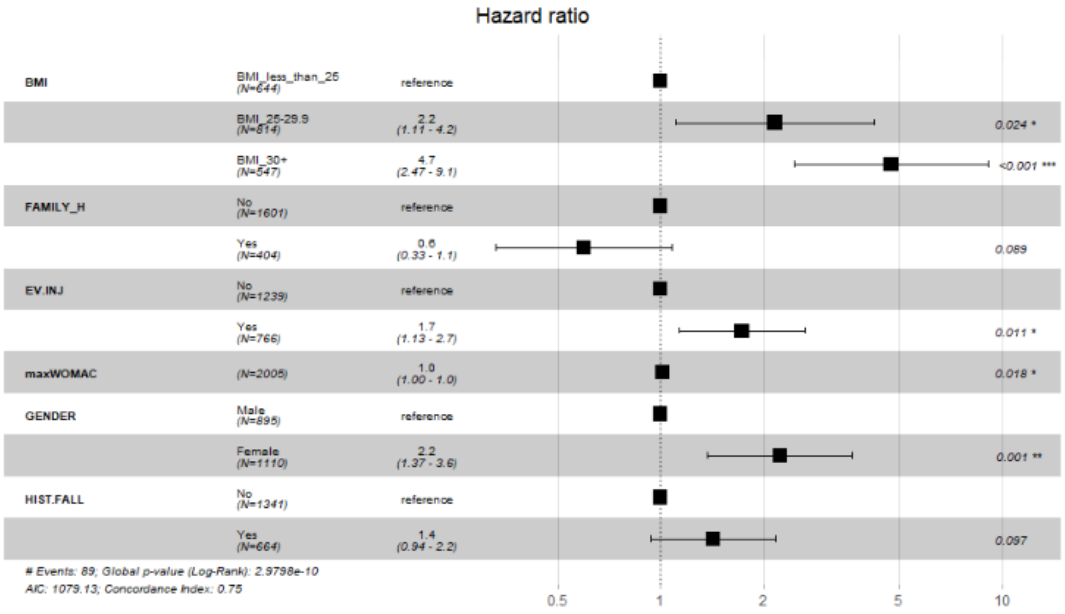


Figure 4-41: Forest plot showing the hazard ratios for the variables included in the backward stepwise Cox model.

To ensure the modelling of the variables is appropriate for the assumptions made about proportional hazards, a test is performed. Based on the results shown in Figure 4-42 all of the covariates, along with the model as a whole, follow the proportional hazards assumption. As the forward and backward stepwise models are the same, the proportional hazards calculated are also the same. Therefore, Figure 4-38 shows the corresponding Schoenfeld test plots.

	chisq	df	p
BMI	5.3627	2	0.068
FAMILY_H	0.3862	1	0.534
EV.INJ	0.0537	1	0.817
maxWOMAC	3.0387	1	0.081
GENDER	0.1446	1	0.704
HIST.FALL	0.1848	1	0.667
GLOBAL	9.7857	7	0.201

Figure 4-42: The results from the test for the proportional hazards assumptions on the covariates from backward stepwise Cox regression. All covariates have p-values greater than 0.05, so none are significant.

The survival curves for the backward stepwise model is the same as those for the forward stepwise model, in Figure 4-39.

4.6.2.2.4. Stepwise with Stratification

As the forward and backward models both give the same variables in for the Cox regression, the model variables are consistent. The variables here are BMI, previous history of knee injury (EV.INJ), gender, family history of knee issues (FAMILY_H), a history of falling (HIST.FALL) and the maximum WOMAC score recorded (maxWOMAC).

The next step is to see if there are groups within the cohort. For example, to determine

if there is a high and low risk group, and to establish what the criteria is for inclusion in each group.

To stratify the group into cohorts the first step is to establish a cut point that gives the biggest separation in the subjects. Figure 4-43 shows the histogram along with the cut point that is used to define the two strata that will produce a high and low risk cohort.

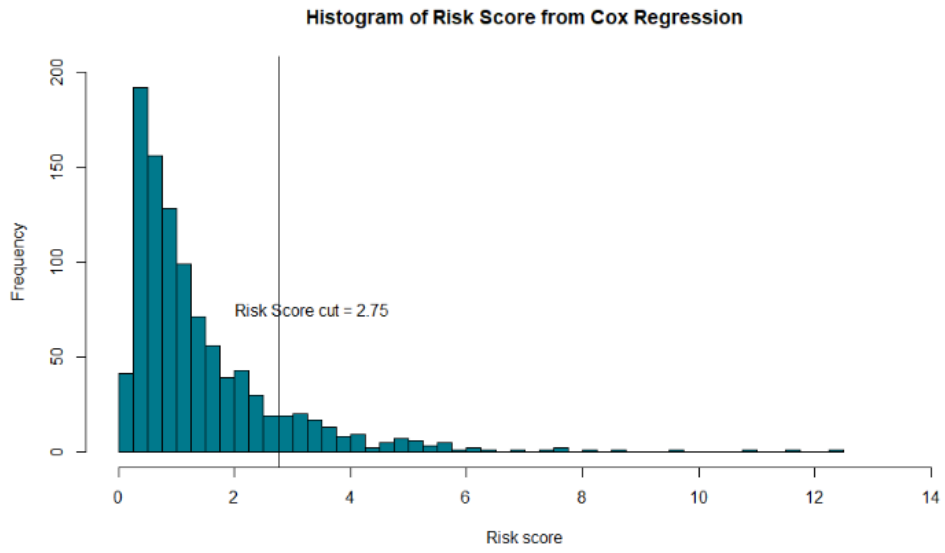


Figure 4-43: Histogram of the risk score for the training data when establishing where to add the cut point. In this case, the cut point falls where the risk score is 2.75.

Figure 4-44 shows the survival curves for the training data after it has been split into the risk stratification cohorts. The left curve shows the actual data and how it is divided. Note, that in cohort 2 the last recorded event occurs within the 5-year span at day 1642. The curve on the right shows the predictions once they have been split into the risk stratification cohorts. As there is no real overlap of the confidence intervals, the two populations are different from each other. Therefore, the idea of two risk groups is valid for this dataset.

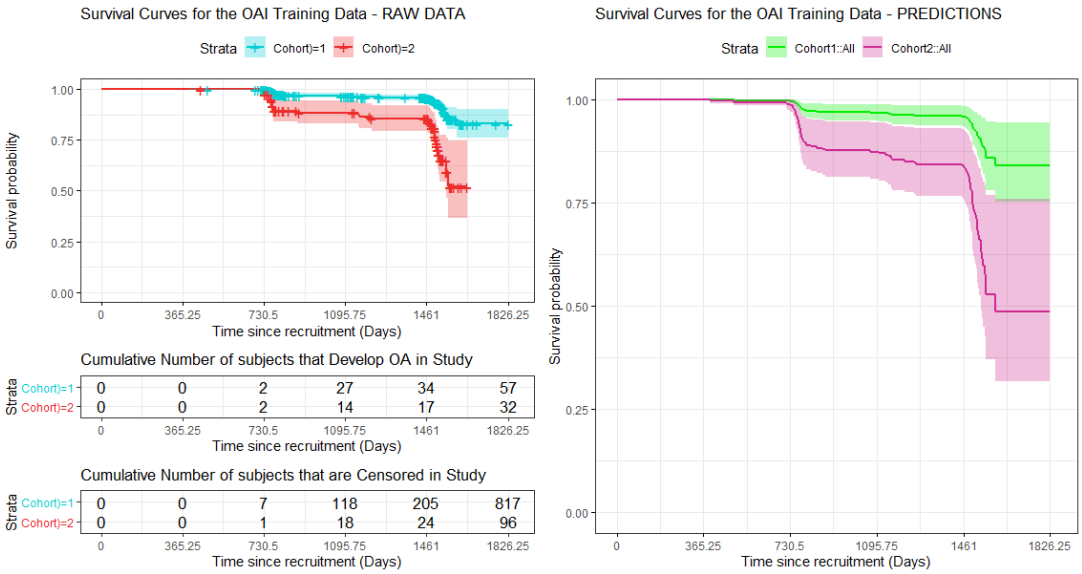


Figure 4-44: Stratification curves on the training data showing the high and low risk cohorts. Note the last even recorded in cohort 2 within the 5-year span is at day 1642. Stratification curves on the training data showing the high and low risk

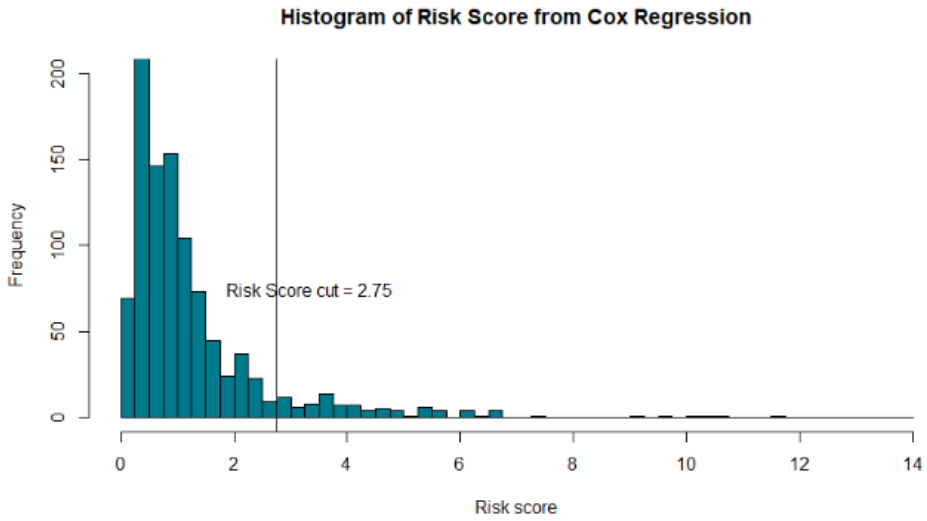


Figure 4-45: Histogram of the risk score for the test data showing where the predetermined cut of the risk score falls.

The histogram in Figure 4-45 is showing the risk scores from the test set and the predetermined cut point and the way this divides the data. Figure 4-46 shows the stratification curves on the test data for both the raw data and the predictions produced on the test set. Once again, like the training data, the stratification curves produced are well separated with no crossover on the confidence intervals.

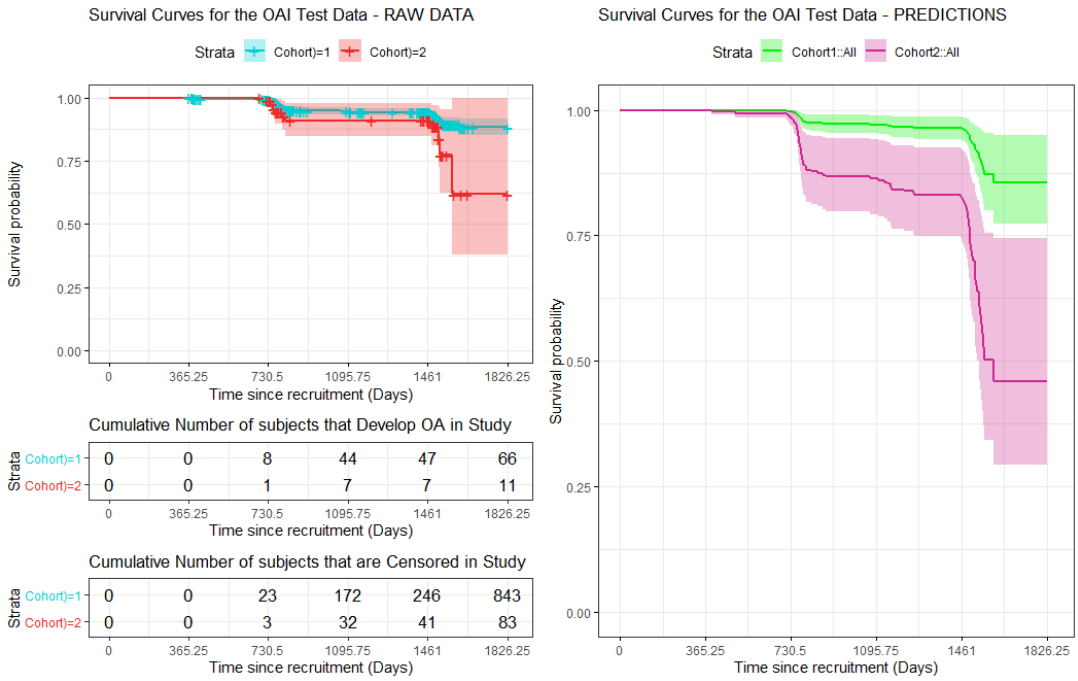


Figure 4-46: Stratification curves on the test data showing the high and low risk cohorts using the predetermined cut points calculated on the training data.

For the model to have clinical value the findings of the two risk cohorts need to be translated into human terms. For example how the features influence that individual in relation to which risk group they will belong. The proportions are shown in Figure 4-47.

The proportion plots are useful as they can be used easier to profile the groups in each cohort. For example, in Cohort 2 all of the subjects are female with a BMI over 25, and they have all had previous knee injuries. The majority of the people in cohort 2 have no family history of knee problems, which could mean that those who are aware of the issues their families had already made changes to their behaviours to help prevent them from developing KOA.

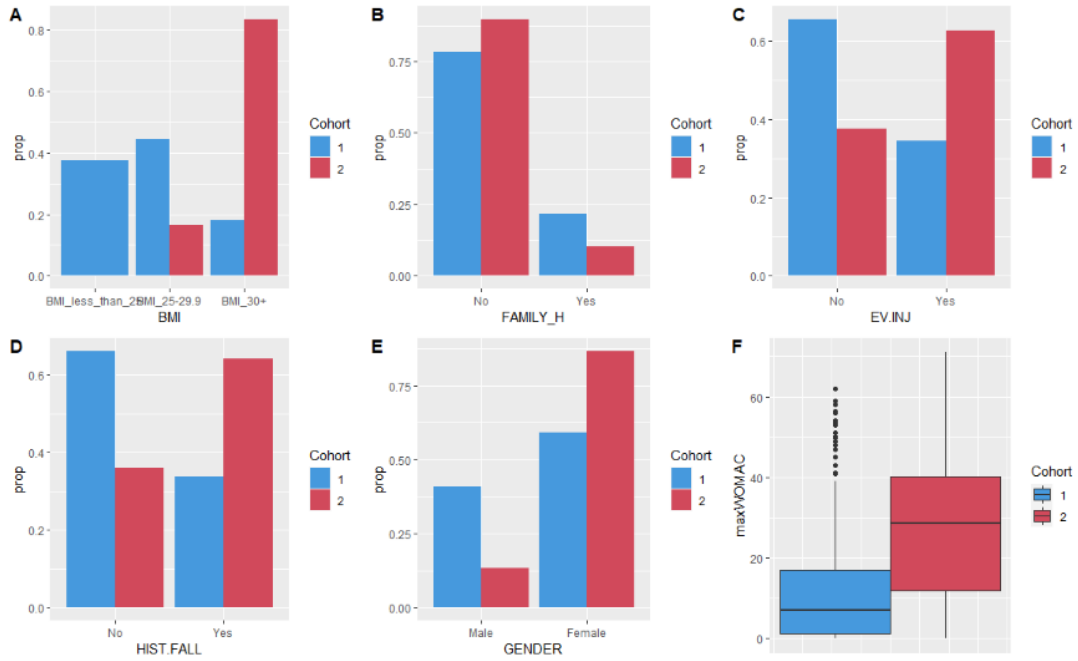


Figure 4-47: The cohort profiles per variable for the different strata. The blue bars show cohort 1 and red show cohort 2. This representation of the profiles is the proportion of the group in each data category per cohort for the training set.

The same plot for the test data is in Figure 4-48. The same patterns that are evident in the training set are also present in the test set.

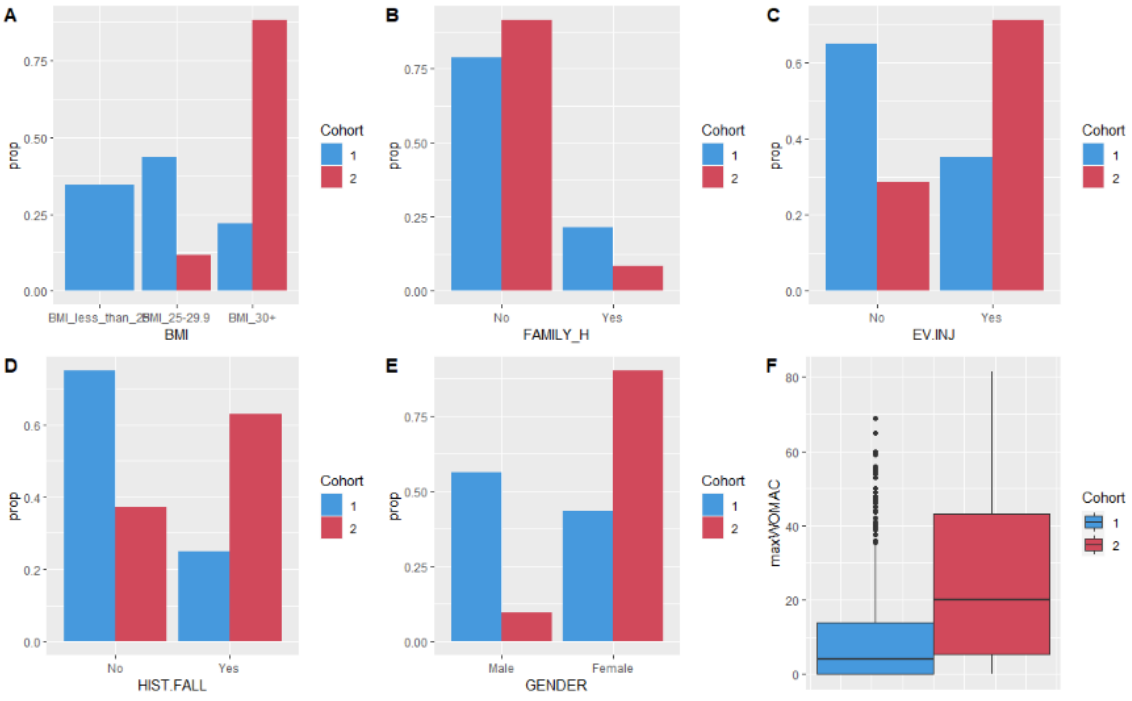


Figure 4-48: The cohort profiles per variable for the different strata. The blue bars show cohort 1 and red show cohort 2. This representation of the profiles is the proportion of the group in each data category per cohort for the test set.

4.7. Exploration of Discrete Time Survival Analysis

4.7.1. Motivation and Justification

Based on the results and the Kaplan-Meier curves shown in this chapter it is evident that there is some periodicity in the study protocol, therefore an analysis using discrete time is presented as the exact time of the onset of OA is unknown so a window of time is considered as the onset period.

It is apparent from the 5 and 7-year analysis that there are steps in the follow-up. These steps mainly occur at two-year intervals with steps naturally occurring at roughly 2.5, 4.5 and 6.5 years after the start of the study. In order to maximise the window for events to be captured, the data used in the discrete time analysis is the 7-year cohort. Following from the natural steps in the data, there are three time division in the 7-year follow up, the start of the trial to 3 years, 3-5 years and 5-7 years. A summary of the events for the data when categorised into three time intervals is shown in Table 4-5.

Table 4-5: Summary of the number of subjects that developed KOA in each time window in the discrete time study.

	No Disease	**Developed KOA**	**Developed since Previous Interval**
End of Interval 1	2003	92	92
End of Interval 2	1929	166	74
End of Interval 3	1839	256	90

The data now has time as a category, either 1, 2 or 3 referring to the time point in the trial the clinical assessment visits took place, 0-3 years, 3-5 years, or 5-7 years. Each data ID referring to a different individual, has three instances within the data, one falling in each time period. This is to model each of the discrete intervals as a logistic regression, using the relationship in the Equation 4-7 to calculate the hazards. The formulas given in Equation 4-10 (1-3) are used to determine the survival at time intervals one, two and three.

$$\begin{aligned}\text{Onset of KOA}(T_i) &= h(x, t = i) \\ &= \text{Family History} + \text{BMI} + \text{Previous Injury to Knee} \\ &+ \text{OA in Any Joint} + \text{Gender} + \text{time1}(T_i) + \text{time2}(T_i)\end{aligned}$$

Equation 4-7: The logistic regression formula that is used to allow the discrete time implementation of the analysis.

As for previous modelling, stratification and cohort profiles are also presented for the two strata created from the model to demonstrate the profiles of those at high and low risk for developing KOA within the timeframe.

4.7.2. Theory

The discrete time Cox model for h_k is given in Equation 4-8. It is calculated as:

$$\log\left(\frac{h_k}{1-h_k}\right) = \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + \beta_{time1} time_1 + \beta_{time2} time_2 + \beta_0$$

Equation 4-8: The formula for the discrete time Cox model.

Where h_k ~ the hazard in time interval k,

$x_1, x_2, \ldots, x_n$ ~ The set of n covariates,

$\beta_1, \beta_2, \ldots, \beta_n$ ~ The coefficients that measure the impact of the covariates,

$time1, time2$ ~ The covariates for time1 and time2,

$\beta_{time1}, \beta_{time2}$ ~ The coefficients that measure the impact of the time1 and time2 covariates

The time intervals T1, T2 and T3 are coded in the data as shown in Table 4-6. The changing of the time code from binary indicators to two columns was to ensure that time is not artificially correlated with other variables, or any such combination. The format for coding, shown in Table 4-6, removes this problem and can be later used when splitting the predictions to calculate the survival.

Table 4-6: An index table showing the way that the time intervals were coded for use in the model.

	Time 1	**Time 2**
T1	0	0
T2	1	0
T3	0	1

$T1, T2$ and $T3$ are time intervals. The information presented in Equation 4-9 is used to separate the predicted probabilities that will be used to calculate the survival at each time point, as given in Equation 4-10.

1. $h(x, t = 1)$ when $P1 = 1$
2. $h(x, t = 2)$ when $P2 = 1$
3. $h(x, t = 3)$ when $P3 = 1$

Equation 4-9: The hazard functions, 1, 2 and 3, calculated at each time point, t, as defined by categorising the time intervals.

1. $S(x, t = 1) = 1 - h(x, t = 1)$
2. $S(x, t = 2) = \big(1 - h(x, t = 1)\big) \times \big(1 - h(x, t = 2)\big)$
3. $S(x, t = 3) = \big(1 - h(x, t = 1)\big) \times \big(1 - h(x, t = 2)\big) \times \big(1 - h(x, t = 3)\big)$

Equation 4-10: The survival equations that are used, using the relationship with the hazard function for survival at time intervals t = one (1), two (2) and three (3).

Modelling the hazard as a logistic regression, as given by Equation 4-7, then means that the predicted probabilities can be used to calculate the survival probability for each time interval. Equation 4-10 shows the relationship with the survival and hazard functions for the different time intervals and how they are used to calculate the survival for that given time-period. For example, for the survival of the third time interval, the hazard function values for time intervals one, two and three are all used, as the final survival value is the cumulative effect of the survival in the other intervals.

For the implementation of the discrete time analysis, the split sample validation that has been used throughout the thesis is used. Initially the logistic regression model produced the beta values that were then used to determine the stratification cut-off points. This step allowed a high and low risk cohort to be developed such as with the Cox Regression analysis. Following from the stratification, the cohort profiles describe the populations that make up the high and low risk groups for developing KOA within the 7 year, 3-window time frame.

4.7.3. Results

A multivariate logistic regression model is used to implement the discrete time survival analysis modelling described in this section of the thesis. This model is used to calculate the hazard odds which is then used to calculate the survival. The multivariate implementation is used to assess how the covariates jointly influence the probability of the subject developing KOA. The significant variables, based on this, that have a strong association with the outcome are given with a $p - value > 0.05$. In this analysis, the significant variables are *BMI, Previous knee injury* and *previous OA in another joint.* As with the

studies detailed earlier in this chapter, BMI and Previous knee injury are among the variables found to be significant. These results are illustrated in Figure 4-49.

```
Call:
glm(formula = outcome ~ FAMILY_H + BMI + EV.INJ + PREV.OA + GENDER +
    time1 + time2, family = binomial(link = "logit"), data = train_l)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.1301  -0.4438  -0.3383  -0.2502   2.9082

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -6.74695    0.38088 -17.714  < 2e-16 ***
FAMILY_HYes  0.03420    0.13424   0.255 0.798898
BMI          0.11599    0.01186   9.776  < 2e-16 ***
EV.INJYes    0.46969    0.11129   4.220 2.44e-05 ***
PREV.OAYes   0.38892    0.11140   3.491 0.000481 ***
GENDERMale  -0.16958    0.11283  -1.503 0.132865
time1        0.70390    0.15947   4.414 1.01e-05 ***
time2        1.19632    0.15062   7.943 1.98e-15 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2636.7  on 4709  degrees of freedom
Residual deviance: 2436.0  on 4702  degrees of freedom
AIC: 2452

Number of Fisher Scoring iterations: 6
```

Figure 4-49: The logistic regression implementation of the model used to calculate the hazard.

The next step in the analysis is to determine if there are groups within the cohort. This is the process in establishing if the data can be split into high and low risk groups and what features are likely to contribute to each.

To identify the groups, the data must first be stratified. Using the stratification technique, discussed in Section 4.4.7, a cut point is identified. This cut point provides the point within the data that gives the biggest separation of the subjects. Figure 4-50 shows the histogram along with the cut point that is used to define the two strata that will produce a high and low risk cohort.

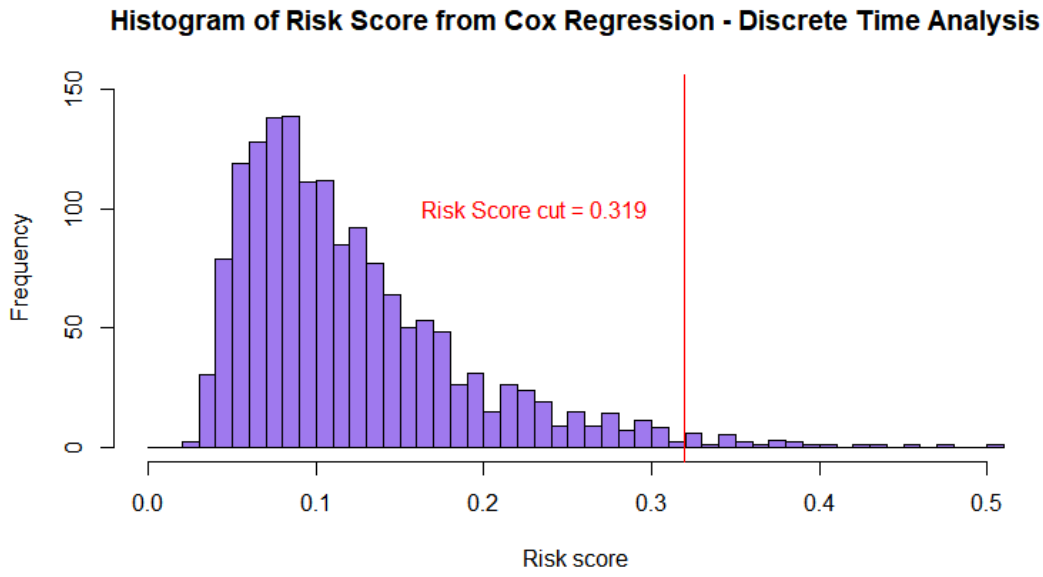


Figure 4-50: The histogram of the risk score for the training data when establishing where to add the cut point. In this case, the cut point falls where the risk score is 0.319.

Figure 4-51 shows the survival curves for the training data after it has been split into the risk stratification cohorts. The thicker lines show the Kaplan-Meier curves whilst the thinner lies and points show the predictions calculated using the discrete time implementation. The prediction curves show separation, and only overlapping with the KM curve meaning that the cohorts are distinct. Figure 4-52 shows the stratification curves on the test data for both the raw data and the predictions produced on the test set. Once again, like the training data, the stratification curves produced are well separated.

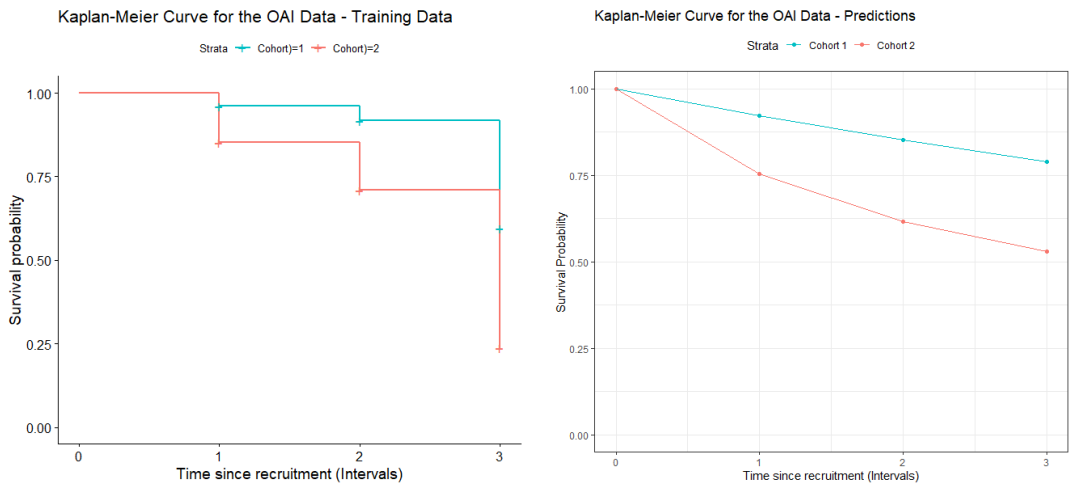


Figure 4-51: The stratification curves on the training data showing the high and low risk cohorts. The plot on the left shows the Kaplan-Meier curves whilst the plot on the right shows the predictions calculated using the discrete time implementation.

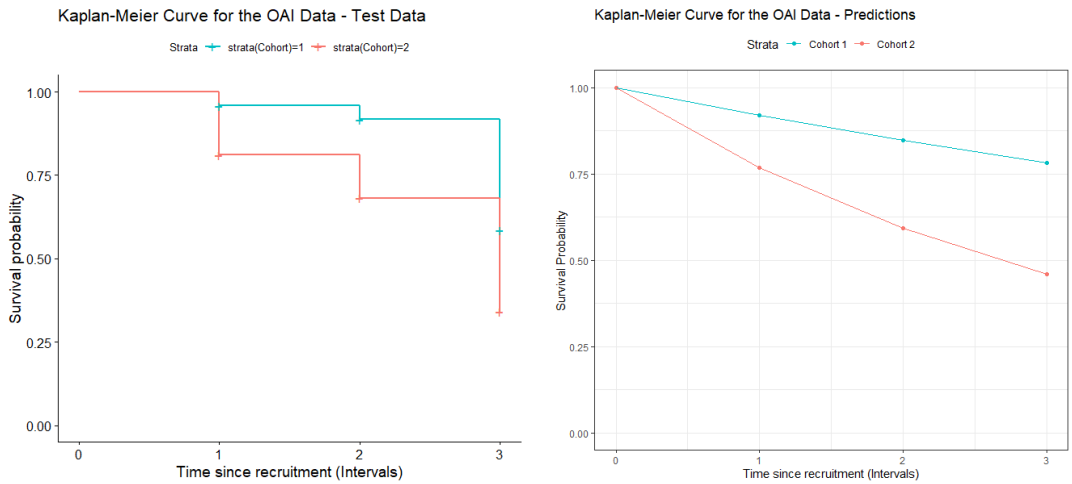


Figure 4-52: The stratification curves on the test data showing the high and low risk cohorts. The plot on the left shows the Kaplan-Meier curves whilst the plot on the right shows the predictions calculated using the discrete time implementation.

For these finding to have meaning the risk cohorts need to be explained in human terms, in relation to the variables in the model. For example how the features influence that individual in relation to which risk group they will belong. The proportions for the training and test sets are shown in Figure 4-53 and Figure 4-54 respectively.

The proportion plots are useful as they can be used easier to profile the groups in each cohort. From the training data cohorts shown in Figure 4-53, the factors that make up the majority of Cohort 1 are a lower BMI, no family history of KOA, no previous knee injuries, no previous diagnosis of OA in any joint in the body and there is a mix of males and females. The features likely to indicate class membership into Cohort 2 are high BMI scores above 30, Previous knee injury, previous diagnosis of OA in another joint in the body and being female. No family history of KOA seems to have a protective effect on the development of KOA but in the group with a family history of the disease, there are slightly more in Cohort 2.

Many of the traits from the training data are also true in the test data, shown in Figure 4 54, with one notable exception. Having no family history of KOA is now a factor more likely to indicate membership to Cohort 2. However, the same patterns that are evident in the training set are also present in the test set.

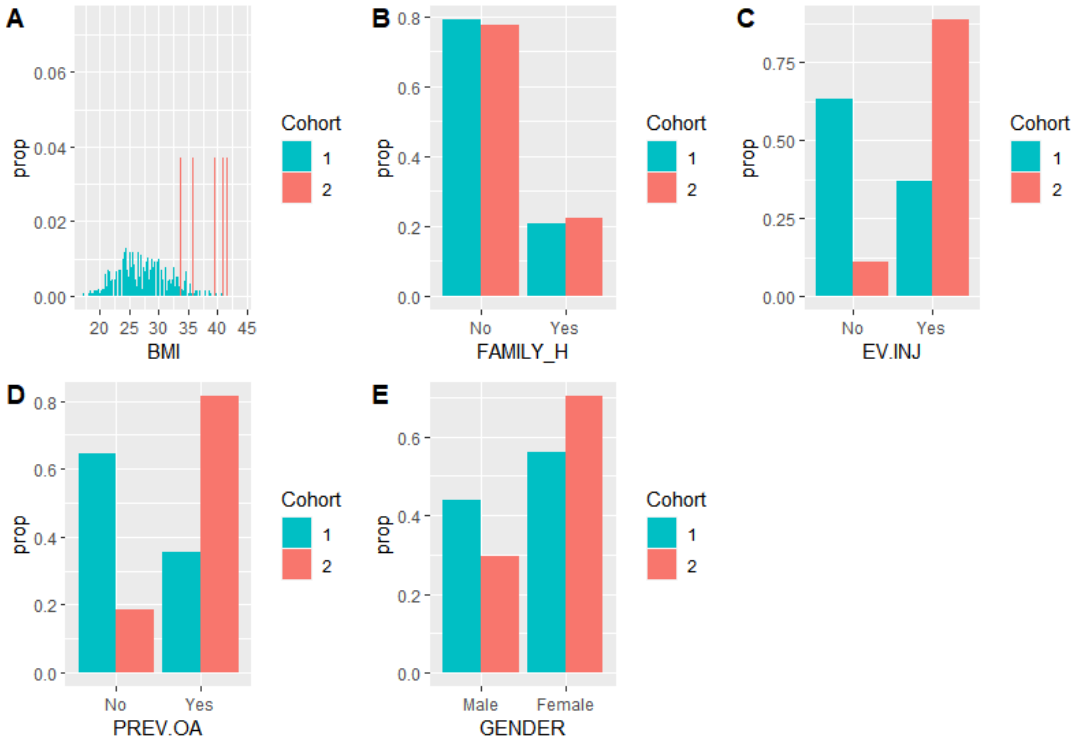


Figure 4-53: The cohort profiles per variable for the different strata. The green bars show cohort 1 and red show cohort 2. This of the profiles is the proportion of the group in each data category per cohort for the training set.

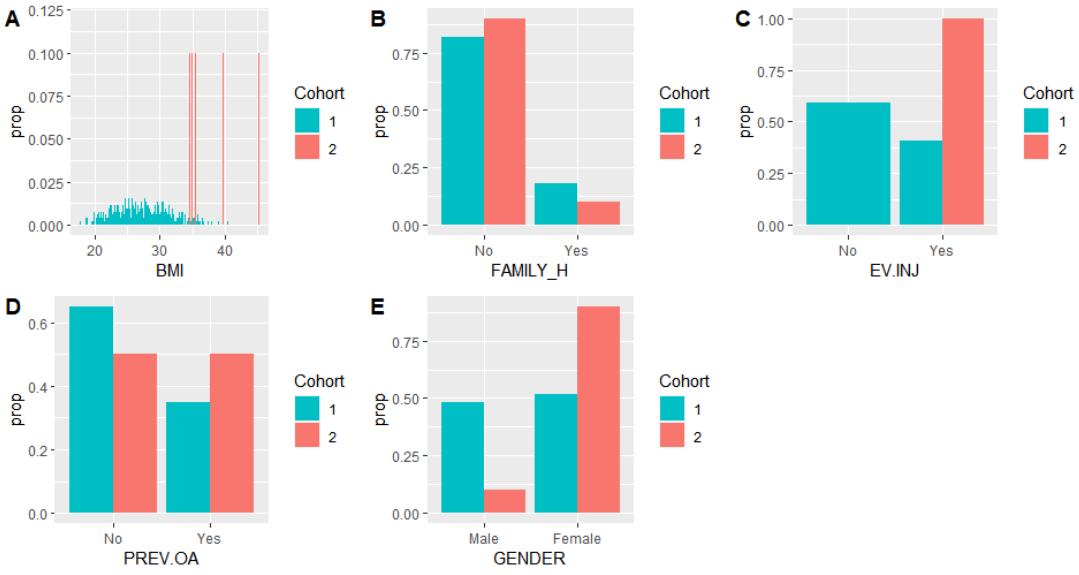


Figure 4-54: The cohort profiles per variable for the different strata. The green bars show cohort 1 and red show cohort 2. This of the profiles is the proportion of the group in each data category per cohort for the test set.

4.8. Discussion

The main point of this chapter was to establish if KOA could be modelled using survival modelling and if this was the case, to see if there was any way to establish risk groups within the population. Both of these things were possible.

Firstly, when considering a follow-up timespan both a seven-year and a five-year window were tracked. Although there are merits to the longer follow-up time, given how the data differed between the training and test sets the slightly smaller observational window was preferred. This was due to the differing in observed survival between the training and test sets in the 7-year analysis, which was still present but at a more acceptable level in the 5-year study. Based on this, any further analysis, such as model validation, will be carried out using the 5-year cohort.

When considering the variables used in each model the five-year had six covariates, whereas the seven-year had five. The variables were selected from the same pool and contained four, which were consistent: BMI, Gender, Family history of knee problems and surgery, and previous knee injury. The extra variable in the seven-year study is does the subject have any previous OA diagnosis in other joints in the body, and the extras in the five-year study are history of falling and WOMAC score.

When looking at the stratification curves generated on the model using risk scores, we see that BMI is useful in defining risk cohorts. BMI is also helpful in the stratification as it gives the option to show how a persons risk of developing KOA may change if the subject can change their weight. This type of tool may be useful in clinical settings such as at the GP level. It could be used for patient education in people who do not currently have KOA but fall into the age bracket where the disease begins to manifest.

An interesting finding is that the age variable is not present when using feature selection in either the five or the seven-year cohorts. This may be due to the other features contributing in a more significant way to the onset of KOA. Similarly, other OA diagnosis only becomes important in the seven-year cohort suggesting that the longer a person suffers with OA in any joint the more at risk they are of developing this is another joint elsewhere in the body.

For the discrete time analysis of the data, the separation of the survival curves due to stratification, for both the training and test data is clear. The better smoothness of the discrete time fits in Figure 4-51 and Figure 4-52, and the differences with the predictions

for continuous time are likely caused by the interval censoring which is better taken into account by the use of discrete time intervals.

The next thing following this will be to validate any of the existing models devised using the OAI data with the MOST data. The MOST data is from a similar study to the OAI study. Similarly, the OActive data is available to use to validate the existing models, where appropriate.

Following on from this step, I will investigate the way risk is different for males and females. To do this I will split the data into male and female, include gender specific variables, where applicable, and repeat the process of variable selection and stratification. By making use of the variable selection, it will be possible to identify which, if any variables that relate to one gender or another are significant to the development of KOA. The aim is that stratification will give a high and low risk group for each gender. If this was the case, there is the potential to find a more clinically useful way to treat male and female subjects, instead of the current blanket approach.

Chapter 5: External Model Validation for Diagnostic and Prognostic Models

The research defined in this chapter has been published in PLOS One.

McCabe PG, Lisboa P, Baltzopoulos V, Olier I.
Externally validated models for first diagnosis and risk of progression of knee osteoarthritis. 2022. PLOS ONE 17(7): e0270652.
Available from: https://doi.org/10.1371/journal.pone.0270652

5.1. Introduction

In Chapter 3 and Chapter 4, different models have been developed for use in clinical settings to assist in clinical decision-making processes. An important step that is required before a model is deployed for clinical use is model validation.

Model validation usually refers to the process in which hyper parameters in a model are tuned to improve the model performance, and ensure the best possible results. The validation of a model is a process that is also used to verify outputs are as expected and therefore confirm the model is robust [137] and it is a vital step in showing that the model is less likely to overfit to new data.

When developing any kind of model with the aim of it being used in industry there is a necessity in ensuring that the model is capable of performing well on the data provided, not only the data used in the modelling process. Validation confirms that the model can be applied to datasets outside of the data scope used in modelling [138]. Having undergone validation, the results and analysis are viewed to be more reliable [139].

Having a validation set to use when developing a model also offers the chance to be able to analyse the model performance during development, and where required, make improvements to the models ability to generalise [137]. Continuing with this point, using a validation set also depicts whether a model has been sufficiently trained so that it can accurately and adequately represent the behaviour of the system, disease or mechanism being modelled and studied [140], [141].

In this instance, and throughout this chapter, validation refers to an independent test set gathered from an external source. The criteria for an external test set is that the data is similar in subject matter but unique from the modelling data. A visual representation of this is shown in Figure 5-1.

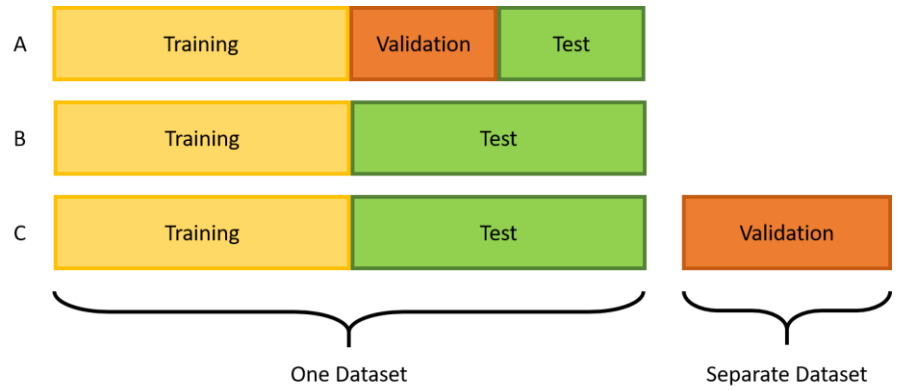


Figure 5-1: Visualisations of different ways to partition a dataset for use in modelling. Configuration A shows the validation set as part of the single dataset used for training, validation, and testing. This validation is used to test the model after it has been trained and the test set would be for final model evaluation. Configuration B shows only a training and test set, where the test set would be used to evaluate the trained model. Configuration C is similar to that of B but makes use of an external validation set. In this instance, validation set refers to independent test set from an external data source. The criteria for an external test set is that it is similar but unique from the modelling data.

When selecting an appropriate data cohort for external validation there are certain perquisites that must first be checked. The independent dataset must come from a population that is comparable to that used for modelling. Comparable data would have the same or similar inclusion and exclusion criteria, and include data with the same predictor and outcome measures [89]. In the case of the MOST data, used as the external validation data in this analysis, the criteria is extremely similar, with the main difference being in the age – OAI has subjects aged between 45-79 whereas MOST has subjects aged 50-79. There are small differences with the predictor variables but appropriate steps to deal with that are discussed in sections 5.2 and 5.3.

The use of an external validation set is extremely important when considering whether to accept a prediction model for widespread use [142]. Generally, a prediction model will have good performance when tested on data from the same population that was used to train the model, but suffer worse performance when applied to new cases [143]. External data can be gathered from either a retrospective study that has already been done, or prospectively, by creating a new study and enrolling subjects for validating the model that requires assessment [61], [144].

In this chapter, the retrospective approach was used, with MOST used as the external data and OAI being used for model training. The advantage of this is both datasets are of a considerable size, but they are also both from the same geographical location [61]. To undergo a geographical validation in this instance a UK based dataset would be required. This would verify if the US demographic could be applied to the UK for the purposes of KOA prediction modelling. Even though the data is from the same location,

the USA, there is still debate around how well data from a different place will fit a model developed on data from somewhere else, given that the hospitals in this case are different for the OAI and the MOST studies [145].

One study produced a 4 year risk model for predicting the risk of symptomatic KOA using logistic regression, but lacked external validation [146]. In many cases related to KOA, finding a suitable external validation dataset is not possible [34], [41], [43], [45], [54], [56], [57], [147], [148]. However, depending on the research question, the opposite can be true [50], [55], [76], [149]–[151].

One of the aims throughout this work was to make a model that was interpretable and easy to use in a clinical environment. In order to do this, interpretable approaches were used in the modelling of the data and web based applications were developed for accessible use. These user interfaces offer the potential for the tool to double as a clinical aid and a resource for patient education. Having the risk model displayed in this way allows the patient to easily see and understand the way in which the factors relating to their life impacts on their individual risk of having or developing KOA. If the tools were to be used as a clinical aid they may help to improve patient flow from query to diagnosis and better allow for more in depth investigations, such as targeted x-rays for those who are on the borderline of having KOA or those at high risk. Another way that the interfaces could serve with clinical use is for patient signposting. Currently, the NHS offer lung screening visits to people aged between 55 and 74 who smoke or have previously smoked [152]. This initiative helps to detect lung conditions, such as cancer, earlier than they would have been picked up, allowing for better disease outcomes and more targeted treatments. Having a similar tool in place for KOA may help identify those at high risk of developing the disease in the next five years and allow for more targeted advice to those individuals, potentially having a positive outcome in relation to delaying the onset of KOA.

The work in this chapter builds on that from Chapter 3 and Chapter 4 by extending the approach to ensure the model can stand up to an external dataset that is independent from the training data. The models in this chapter are very similar to those in the previous chapters, but may have been subject to minor modifications to be able to validate the models with the use of the MOST and OActive datasets.

The objective in this chapter is to develop and validate a diagnostic and prognostic model to determine the presence of KOA at baseline or to calculate if a subject is at high or low

risk of developing KOA in the next five years. This work extends the interpretability of the models, which can then be converted into web apps that have the potential to be used in clinical settings, such as GP surgeries to help streamline both diagnosis and patient education, leading to better clinical management and self-perceived quality of life for those with KOA.

Chapter aims

- Demonstrate the model performance for both the diagnostic and survival models on suitable external datasets.
- Develop web applications for the usability of the validated models in clinical practice.

5.2. Specifics of the data used in chapter

In this chapter, data from OAI is used for training and testing, whilst data from MOST and OActive are used for the external validation.

5.2.1. Class Definition

Clinical KOA in this analysis, as throughout the thesis, is a binary outcome defined by the KL score. Scores zero and one are classified as no clinical KOA, and therefore zero. A KL score of two or above determines the positive class, clinical KOA, therefore classified as one as the binary indicator. These KL grades have been determined by a clinician from analysing the X-rays taken as part of the study. This definition is consistent across the OAI, MOST and OActive datasets.

5.2.2. OAI

For the diagnostic modelling cohort, the sample, including all of the missing values had a size of 4796. After removing the subject who have no KL grade leaves a sample size of 4507. Finally, removing those subjects who have missing values in any portion of the variable sets leaves a usable cohort of 2707 subjects in the complete case analysis. The OAI data training and test sets have a prevalence of KOA at 40% and 39% respectively. The rationale behind a split sample approach was discussed in section 2.7. The summary statistics for the diagnostic modelling data are shown in Table 5-1.

Table 5-1: Diagnostic model summary information of the Osteoarthritis Initiative (OAI). The OAI training data was used to develop the models. The variables are listed with the different options each can take.

Diagnostic Model Data			
Variable	**OAI**		
	Total	**Training Set**	**Test Set**

		N = 2707	*n = 1353*	*n = 1354*
BMI	Less than or equal to 25	754	383	371
	More than 25	1953	970	983
Baseline Symptoms	No	2042	1028	1014
	Yes	665	325	340
KPACT30	No	2068	1031	1037
	Yes	639	322	317
Knee swelling	No	1970	993	977
	Yes	737	360	377
Gender	Male	1250	622	628
	Female	1457	731	726
Difficulty Upstairs	No	1352	660	692
	Yes	1355	693	662
Age	45 – 50 years	373	184	189
	50 – 55 years	572	281	291
	55 – 60 years	457	232	225
	60 – 65 years	402	195	207
	65 years or over	903	461	442
Knee Stiffness in the past 30 days	0 days	2069	1032	1037
	1 – 7 days	289	152	137
	8 – 14 days	118	59	59
	15 – 21 days	114	58	56
	22 days or more	117	52	65
KL status	KL < 2	1627	806	821
	KL 2+	1080	547	533

For the prognostic modelling, only subjects with no baseline KOA and at least one follow up measurement could be included, as this approach considers the time to change state from no disease to active KOA. Removing any subjects that had KOA at the baseline

assessment and did not meet the follow-up filter leaves a sample of 2314 subjects. These subjects had no OA, in other words, a KL score of 0 or 1 at baseline. Considering basic demographic features for subjects where there are no missing values, the usable subject cohort is comprised of 2136 subjects. Filtering out any subjects with the event of interest outside of the 5 year cut off results in a sample size of 2005 subjects. The variables for the prognostic model, along with the summary statistics are shown in Table 5-2.

Table 5-2: Prognostic model summary information of the Osteoarthritis Initiative (OAI). The OAI training data was used to develop the models, while the test set provided a level of internal performance evaluation. The number in round brackets on the WOMAC row is the median value for that variable.

PROGNOSTIC MODEL DATA				
VARIABLE		**OAI**		
		Total *N = 2005*	**Training Set** *n = 1002*	**Test Set** *n = 1003*
BMI	Less than 25	644	329	315
	25 – 29.9	814	409	405
	30+	547	264	283
FAMILY HISTORY	No	1601	800	801
	Yes	404	202	202
EVER INJURED KNEE	No	1239	621	618
	Yes	766	381	385
HISTORY OF FALLING	No	1341	624	717
	Yes	664	378	286
GENDER	Male	895	373	522
	Female	1110	629	481
WOMAC		0-82 (6.8)	0-71 (8)	0-82 (5)
KOA	Censored	1839	913	926
	Develop KOA	166	89	77

5.2.3. MOST

The data from MOST for the diagnostic model validation was prepared in the same way as for the OAI data. All subjects were required to have no missing values for the variables present and an initial KL grade from the baseline assessment. The only difference for the MOST data is that one variable, knee_swell, was not present in the dataset. Therefore, to establish predictions from this data including this covariate, we marginalised over the other variable combinations and produced predictions. The MOST validation set prevalence is at 60%, which is higher than that used to train the model. The summary statistics for the diagnostic model validation data are shown in Table 5-3.

Table 5-3: Diagnostic model summary information of the Multicentre Osteoarthritis Study (MOST) datasets. The MOST data was used as external validation. The variables are listed with the different options each can take. As the knee swelling data is missing in the MOST dataset, the predictions are marginalised over the OAI data to find outcomes that match the cases for the other variable combinations. The number in brackets represents the number per variable after marginalisation has taken place.

Diagnostic Model Data		
Variable		**MOST** *N = 2006* *After Marginalisation n = 831*
BMI	Less than or equal to 25	792 (68)
	More than 25	1214 (763)
Baseline Symptoms	No	439 (139)
	Yes	1567 (692)
KPACT30	No	167 (142)
	Yes	1839 (689)
Knee swelling	No	*NA*
	Yes	*NA*
Gender	Male	742 (298)
	Female	1264 (533)
Difficulty Upstairs	No	136 (30)
	Yes	1870 (801)
Age	45 – 50 years	108 (45)
	50 – 55 years	416 (198)
	55 – 60 years	350 (151)
	60 – 65 years	390 (166)
	65 years or over	742 (271)
Knee Stiffness in the past 30 days	0 days	1242 (144)
	1 – 7 days	266 (251)
	8 – 14 days	76 (55)
	15 – 21 days	104 (91)

	22 days or more	318 (290)
KL status	KL < 2	792 (272)
	KL 2+	1214 (559)

In the same way as for diagnostic model, the prognostic model was validated using the MOST dataset. Similarly, to the data for the diagnostic cohort, the prognostic variable subset contained missing values. The missing values were only present in three variables. Therefore to work with the MOST data, and retain a cohort of a meaningful size, imputation was required. To impute the data, filling in the missing values to ensure a sufficiently sized dataset mean imputation based on the training data was used. For the imputation in the MOST data, the family history imputation is 'No', along with the history of falling, and the imputation for the WOMAC score given as a value of 8, the mean of the group. The rationale behind the imputation approach is discussed in Chapter 2, section 2.5, and later in this chapter, in section 5.3. The summary of the prognostic data, along with the effect of the imputation are shown in Table 5-4. The amounts of NA values are shown in square brackets next to the value they are imputed to. The family history imputation is 'No', along with the history of falling. For family history, there are 352 cases where the imputed value is recorded and history of falling has 1003 cases imputed. The imputation for the WOMAC score is shown in italics, with a value of 8 and this was used in four cases. The number in round brackets on the WOMAC row is the median value for that variable, which is 10.

Table 5-4: Prognostic model summary information of the Multicentre Osteoarthritis Study (MOST) dataset. The MOST data was used as external validation. The variables are listed with the different options each can take. To ensure a usable sized dataset NA values are present in the MOST cohort. The amounts of NA values are shown in square brackets next to the value they are imputed to. The family history imputation is 'No', along with the history of falling. The imputation for the WOMAC score is shown in italics, with a value of 8. The number in round brackets on the WOMAC row is the median value for that variable.

Prognostic Model Data		
Variable		**MOST** *N = 1155*
BMI	Less than 25	234
	25 – 29.9	478
	30+	443
Family History	No	328 [352]

	Yes	475
Ever Injured Knee	No	671
	Yes	484
History of Falling	No	130 [1003]
	Yes	22
Gender	Male	693
	Female	462
WOMAC		0-82 [4 *8*] Median:(10)
KOA	Censored	1004
	Develop KOA	151

5.2.4. OActive

The data from OActive can only be used on the diagnostic model validation as there are no follow-ups after the initial visit. All subjects were required to have no missing values for the variables present and an initial KL grade from the baseline assessment. As the model was built using data from the OAI dataset, it includes variables that are not present in the OActive data. To use the data for predictions, we marginalised over the other variable combinations and produced predictions based on these. The common variables are shown in bold in Table 5-5.

Table 5-5: Variables in the OAI risk model for Propensity of Presenting, with those in bold highlighting the variables that are not present in the OActive dataset.

Diagnostic Model Data		
Variable		**OActive** *N = 233* *Useable cohort n = 206*
BMI	Less than or equal to 25	40
	More than 25	166
Baseline Symptoms	No	*NA*
	Yes	*NA*

KPACT30	No	*NA*
	Yes	*NA*
Knee swelling	No	135
	Yes	71
Gender	Male	60
	Female	146
Difficulty Upstairs	No	*NA*
	Yes	*NA*
Age	45 – 50 years	30
	50 – 55 years	29
	55 – 60 years	29
	60 – 65 years	27
	65 years or over	31
Knee Stiffness in the past 30 days	0 days	*NA*
	1 – 7 days	*NA*
	8 – 14 days	*NA*
	15 – 21 days	*NA*
	22 days or more	*NA*
KL status	KL < 2	72
	KL 2+	134

The variables missing were marginalised. To do this, each subject in the OActive data set was categorised into one of the 8 combinations of the binary variables in the model (Knee_swell, Gender, and BMI overweight) and 5 Age bands, 40 possibilities in all, and the predictions of the OAI model were averaged over the training data filtered into the same combination.

The mean model predictions for the specific values of the four variables present in each row in the data were then used to calculate the AUROC for the OActive data (n=206).

The OActive data cohort is created as described in Figure 5-2.

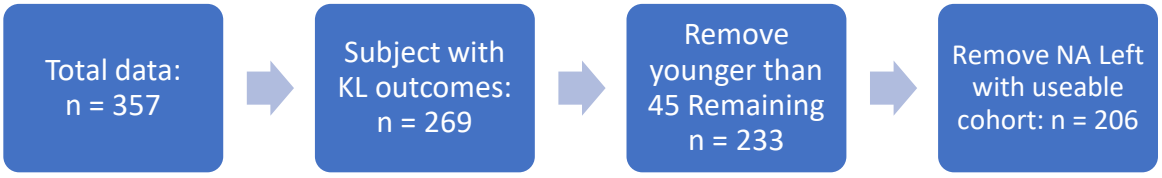


Figure 5-2: Visualisation of data cohort creation.

5.2.5. Pre-Processing

The type of data used in this analysis combines clinical factors, demographic features, self-reported symptoms, and self-reported physical activity data. The clinical and demographic variables include the age, gender, and BMI of the individual, along with information of family history and previous injuries to the knee. The self-reported data set comprises subject's answers to questionnaires relating to their symptoms and how they are impacted, recorded at the first presentation meeting. In a similar approach to the self-reported features, the self-reported physical activity data set consists of answers on questions about how much exercise they take and how this impacts them.

For several features in the original data, more than one column is relevant. To streamline the analysis, and future usability in a clinical setting, we have taken the approach of defining new variables that incorporate the existing ones in a single feature. One such example is for the created variable knee_stiff_day_limit. This looks at how many days in the past 30 a subject has experienced knee stiffness severe enough to limit activity. Several original variables looked at various activities individually, so this approach removes repetition by taking the most severe measure for a subject across all variants of activity. In this situation, if a single variant contains a missing value, the present values are the only ones considered. If all are missing, the consolidated variable is also recorded as missing.

The cohort considered in this analysis was only subjects without any missing values for the selected variable set. This is a complete case analysis [153]. In the preliminary steps of the analysis, not detailed in this chapter, a complete case and imputed analysis were used and compared.

5.3. Study Design

To validate and ensure that these models were not overfitting to the OAI data used to train them, we used the MOST and OActive data to validate the results.

The MOST data was collected from different centres than those used in the OAI study so this helps to determine if the model is able to avoid institutional bias. This helps to

assess if the model can be used outside of the bounds from which the data was collected, and also contributes to showing that a prediction model is more suitable for use in clinical practice [29]. The validation set for the diagnostic model before marginalisation is 2006 subjects, whilst the validation set for the prognostic model is 1155 subjects.

The OActive data was collected from three centres across Europe. Each centre focused on a different subject group. This information is detailed in Chapter 2, section 2.4. The OActive data is only suitable for diagnostic validation, and has a useable cohort size of 233 subjects for model validation.

For the diagnostic model, the variable knee_swell was missing from the MOST dataset, and knee_stiff_day_limit, Diff_upstr, P01KPACT30 and B.Line_Symp were missing from OActive. To combat this we marginalised over the existing variable combinations and produced predictions based on those from the OAI training data. After doing this, some samples were lost due to incomplete matching where the variable combinations in the MOST and OActive data did not have a corresponding combination in the OAI data that was used in the marginalisation. As a result, the sample size is reduced from 2006 to 831 subjects in the MOST data and from 233 to 206 subjects in the OActive data. A large sample size helps to decrease uncertainty and increase precision. Sample size is crucial in ensuring quality research. Large sample sizes allow for better averaging of values and helps to avoid errors that come with small samples. A standard sample size in many domains is less than 100, so data larger than this there are higher levels of replicability, therefore although the sample size has been reduced it is still big enough to have the benefits of a large sample size. This process is described in Algorithm 1.

ALGORITHM 1: USING MARGINALISATION TO RETAIN SAMPLE SIZE FOR THE MOST AND OACTIVE VALIDATION COHORT

Preliminary Step: Train the model using the OAI complete case training data and store the predictions

Input: OAI Training data predictions

Output: A usable marginalised dataset with corresponding predictions

For each categorical variable combination

    Filter corresponding predictions from the OAI training predictions

    Calculate the mean value for the filtered predictions, *train_mean*

    In all rows that meet the conditions input *train_mean* as the corresponding prediction

8 **End**

9 Remove any instance with no predicted value due to not meeting the criteria for marginalisation, as they had no corresponding cases.

Algorithm 1: The pseudocode describing how the marginalisation of the MOST and OActive data took place using the OAI data to calculate predictions.

Marginalisation is a good method to use to deal with missing values in this case as it allows us to make use of data that would otherwise not have been suitable due to missing variables from the whole cohort in the data. The approach works by imputing the missing values for cases where a matching set of features is available with a prediction from the training data that is then identified by case matching corresponding variable combinations. This method assumes that the population of both the data used for imputing, in this case the OAI training data, and the data being imputed, either the MOST or the OActive data, have the same population structure, which could introduce bias into the model. However, as complete case analysis has been used throughout this thesis, marginalisation is good as this approach only imputed for values in which the whole variable contains missing information, for example knee_swell in the MOST dataset. This allows the use of an otherwise unusable dataset for model validation.

For the prognostic model validation, again those with KOA at baseline were removed from the analysis. Date filtering also removed subjects whose event fell outside of the 5-year cut-off. Imputation of missing values with mean predictions calculated from the training set resulted in a sample size of 1155 subjects.

Validating the models with the use of an external dataset further helps to verify that the results have the potential to be adapted for use in clinical decision making processes, as the models are not biased to the training data [142].

5.3.1. Diagnostic Model

The model used in the diagnostic analysis was logistic regression. The logistic regression approach was chosen after considering alternative analysis methods [78]. This method is preferred by clinicians as it mimics their own decision making process.

The goal for the logistic model is to determine, based on eight features relating to a subject, whether they are likely to have KOA and therefore require further investigations into their symptoms. The presence of clinical KOA, KL grade 2 or above is the outcome. The model was trained and tested using the OAI data with 1353 and 1354 subjects respectively.

5.3.2. Prognostic Model

The prognostic analysis uses Cox regression to model how the covariates jointly influences the probability of the subject developing KOA. After modelling with Cox, we created cohorts by risk stratification, to highlight the criteria for being at low and high risk for developing KOA in 5 years from the baseline assessment. To stratify the group into cohorts the first step is to establish a cut-point that provides the largest separation between subjects. The stratification method is explained in Chapter 4, section 4.4.7. This is done with a model containing five variables from the subjects' initial assessment.

The groups used to model this analysis are taken from the original OAI data but removing subjects with KOA at their initial assessment. The model was then trained and tested on 1002 and 1003 subjects from the OAI dataset respectively.

5.3.3. Experimental Set-Up

Logistic regression and Cox regression were optimised with AIC calculated from the test data. All data pre-processing, analysis and subsequent app construction were implemented in R. The logistic regression model uses the built in functions for the analysis in base R. For the prognostic modelling the packages used are survival [154], survAUC [155] and survminer [156]. The example web-based application was implemented with the shiny [157] package, a Web application framework for R.

5.3.4. Measure of Performance

The receiver operating characteristic curve (ROC curve) is a plot that graphically indicates the ability of a model to correctly classify binary outcomes as a threshold is altered. The area under the curve (AUC) is equal to the probability that a classifier will rank a random positive instance higher than a randomly chosen negative one [98]. In the AUC a value of 0.5 indicates a guess, with greater than this being deemed better than a guess, and lower than 0.5 being worse than a guess. The AUC and confidence intervals are calculated using the package pROC [158]. The AUC is calculated using the trapezoidal rule and the 95% confidence interval using 2000 stratified bootstrap replicates.

Sensitivity, specificity, and positive predictive value (PPV) are all statistical measures of the performance of binary classification tests. The sensitivity measures the proportion of actual positives that are correctly identified. The specificity measures the proportion of actual negatives correctly identified. The PPV measure looks at the amount of correctly classified subjects out of the whole group of disease class predictions. In this analysis, these measures are calculated with the caret package [159].

5.4. Results from Analysis

5.4.1. OActive Validation on OAI Model

For the OActive validation, summary statistics for the diagnostic model are listed in Table 5-6, and the ROC curves are in Figure 5-3. The table also includes the information from the OAI training and test results. The results from the OAI data, both training and test, are consistent with the model from Chapter 3. The OActive validation data suggests that the model performs well using this data, with a high AUC and specificity. The low sensitivity indicates that for the OActive data a large amount of positive cases are missed. The validation results have high AUC due to the high level of missingness and the use of marginalisation from the OAI predictions, there is likely bias in the data resulting in such a high AUC.

Table 5-6: Summary statistics for the diagnostic model.

Measure	OAI – Training	OAI – Test	OActive Validation
Sensitivity	0.4790	0.5197	0.4925
Specificity	0.8511	0.8490	1.0000
PPV	0.7629	0.7748	1.0000
AUC	0.7415	0.7475	0.9273
(CI)	(0.7146-0.7683)	(0.7209-0.7742)	(0.8938-0.9608)

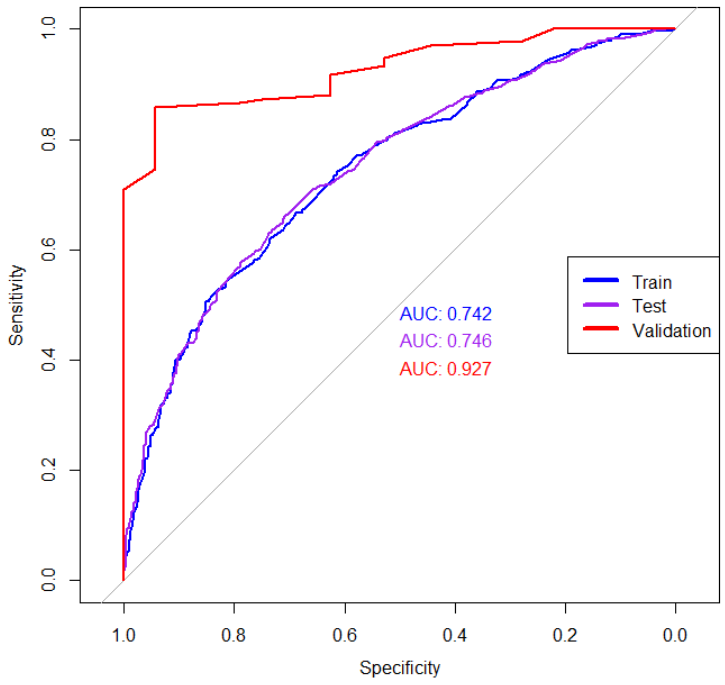


Figure 5-3: ROC curves for the OAI training and test models, and the OActive validation model. The AUC for each curve is listed on the curve.

The results from the OActive validation suggest that this data is a good fit for the model. However, it is important to consider that only 4 of the model variables are present in the OActive dataset, leading to the need for marginalisation over the OAI predictions in order to be able to fit the OActive data to the model. This could lead to bias being fed into the data, due to the small number of predictors in the model because the OActive data only has four of the original variables, resulting in such a high AUC.

In the OActive data, there is a high number of subjects who have been misclassified, as there are 68 cases when a subject with KOA is predicted to not have the disease out of the sample of 206 subjects. There is a level of misclassification in the OAI data also, with 388 subjects of the 1354 receiving a prediction that does not match with the diagnosis. Of the 388 misclassified cases, 261 are deemed to have no KOA when they do, in fact, have KOA. Looking closer at the OAI test data gives an insight at the way the model predicts, and using the test data, it is possible to consider how all covariates, even those not present in the OActive data contribute to the final prediction. The information shown in Table 5-7 considers only the cohort that have been misclassified as no KOA when there is evidence via x-ray that the subject has KOA.

Table 5-7: A table detailing the way the 261 OAI subjects were misclassified and the overall impact this had on the model when applied to the OActive data.

Cohort	*Insight*
143/261	No issue getting upstairs
212/261	No reported knee swelling in last 7 days
257/261	No pain/swelling/stiffness on the day of the baseline assessment (baseline symptoms)
222/261	Made no changes to their activity in the last 30 days as a result of knee pain
222/261	Not had to limit their daily activity due to knee stiffness
192/261	Have a BMI 25+
140/261	Are women.

Based on this finding, it is possible that the discrepancy between diagnosis and prediction when using the OActive data falls in the group of symptomatic vs radiographic knee osteoarthritis. Symptomatic OA is when a person experiences symptoms, such as joint pain, aching and stiffness. Radiographic OA is found by observing features on an x-ray that suggest OA development. It is possible to have symptomatic OA without radiographic OA and vice versa. Up to 60% of people with radiographic KOA may not complain of symptoms [160]. Sometimes, the lack of symptoms are backed up with less severe radiographic OA.

As the OActive data does not contain all of the same variables as the OAI data, several assumptions relating to the predictions are made. By only considering gender, BMI, age and the presence of knee swelling, other symptoms can be left out entirely from the prediction model.

The only symptomatic variable present in the subset of the OActive data is *knee_swell*. This is self-reported and asks if the subject has experienced knee swelling in the last 7 days. Of the 68 subjects that have been misclassified from the OActive data, 63 subjects presented with no knee swelling in the previous 7 days. This may be an influential factor in the model determining between KOA and no KOA. It is worth noting that the variables that are not present in the OActive subset might make the difference in changing cases that were misclassified to class 1, as they hold predictive information about the subjects. Because of this, one limitation using the OActive data is that the model can identify symptomatic KOA, but not radiographic KOA.

5.4.2. MOST Validation on OAI Model

Summary statistics for the diagnostic model, for training, test and validation are listed in Table 5-8, and the ROC curves are in Figure 5-4. The interesting results are that the MOST data performs quite well on a model developed using the OAI data. The MOST results have a high sensitivity, meaning that the model identifies about 90% of all KOA cases.

Table 5-8: Summary statistics for the diagnostic model.

Measure	*OAI – Training*	*OAI – Test*	*MOST Validation*
Sensitivity	0.4790	0.5197	0.9052
Specificity	0.8511	0.8490	0.2353
PPV	0.7629	0.7748	0.5421
AUC	0.7415	0.7475	0.6697
(CI)	(0.7146-0.7683)	(0.7209-0.7742)	(0.6311-0.7082)

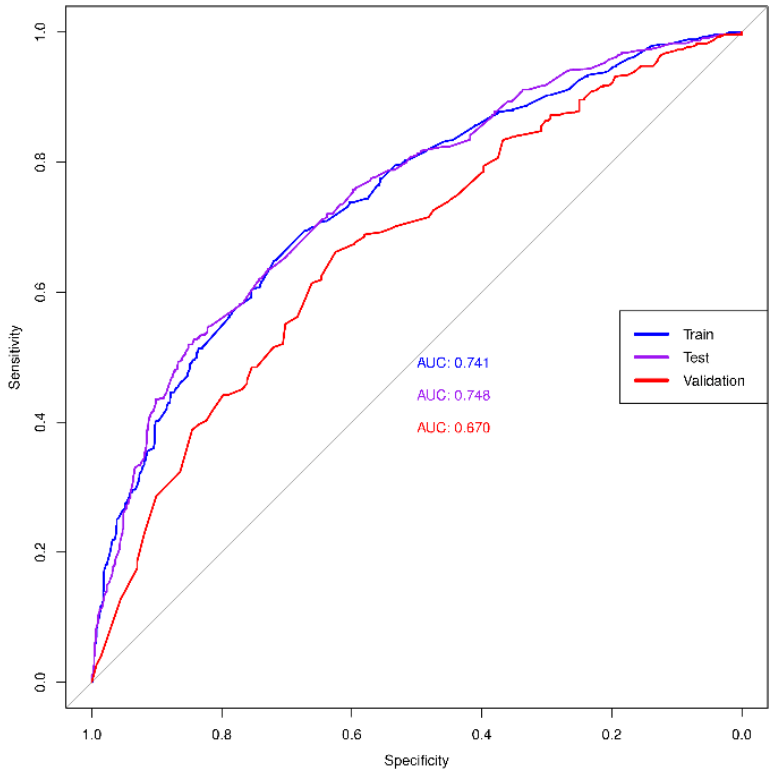


Figure 5-4: ROC curves for the OAI training and test models, and the MOST validation model. The AUC for each curve is listed on the curve.

The odds ratios, modelled on the OAI data, described in Table 5-9, are against the reference category for each variable. These are namely B.Line_SYMP (no knee pain exhibited on the day of the baseline assessment), AGE_bins (participants aged 45 – 50),

BMI_bins (any subject whose BMI was 25 and over), Gender (male participants), Diff_Upstr (no difficulty in getting up stairs), KPACT30 (not modifying activity from knee pain in the past 30 days) and Knee_Stiff (0 days of stiffness in the past 30). Within Table 5-9 the odds ratios for knee_stiff are close to zero, meaning that the variable does not contribute significantly to the outcome. The large confidence intervals show that the variable is not a significant contributing factor to the likelihood of having KOA at the point of first presentation. This could be, in part, due to the loss of data following marginalisation and the removal of cases without matching covariates from the OAI training data. The variable missing from the MOST data was *diff_upstr* but due to incomplete case matching the removal of subjects resulted in a loss of information from the Knee_stiff variable.

Table 5-9: Coefficients of Logistic Regression

	Odds Ratio	95% CI Lower Bound	95% CI upper Bound
Intercept	0.259484	-0.14104	0.66001
Age 50 – 55	1.108447	0.670269	1.546625
Age 55 – 60	1.628192	1.181567	2.074817
Age 60 – 65	2.011177	1.550577	2.471777
Age 65+	2.206814	1.80766	2.605968
BMI BMI less than 25	0.505044	0.220021	0.790067
B.line_symp Yes	4.796757	4.493741	5.099773
Gender female	1.317505	1.069937	1.565073
kpact30 yes	> 100	> 100	> 100
diff_upstr yes	1.097473	0.838753	1.356193
Knee_stiff 1 – 7 days of stiffness	~0.00[a]	-636.498	636.4978
Knee_stiff 8 – 14 days of stiffness	~0.00[a]	-636.498	636.498
Knee_stiff 15 – 21 days of stiffness	~0.00[a]	-636.498	636.498
Knee_stiff 21+ days of stiffness	~0.00[b]	-636.498	636.498

[a] The values for are 0.000001, therefore approximate to 0.00 to 2 decimal places

[b] The values for are 0.000002, therefore approximate to 0.00 to 2 decimal places

As the data for the prognostic model differs from that in the diagnostic model, the training and test sets are also different. The training set and test set comprise of 1002 and 1003 subjects respectively. The external validation set contains n = 1155 subjects. The Kaplan-Meier curve in Figure 5-5 shows lines that represent the full cohort, training, test, and MOST validation data. This shows that the training and test samples are a reflection of the whole cohort, including that used for the validation as when shown with

confidence intervals, they overlap, meaning that the difference between the OAI and MOST datasets are not statistically significantly different.

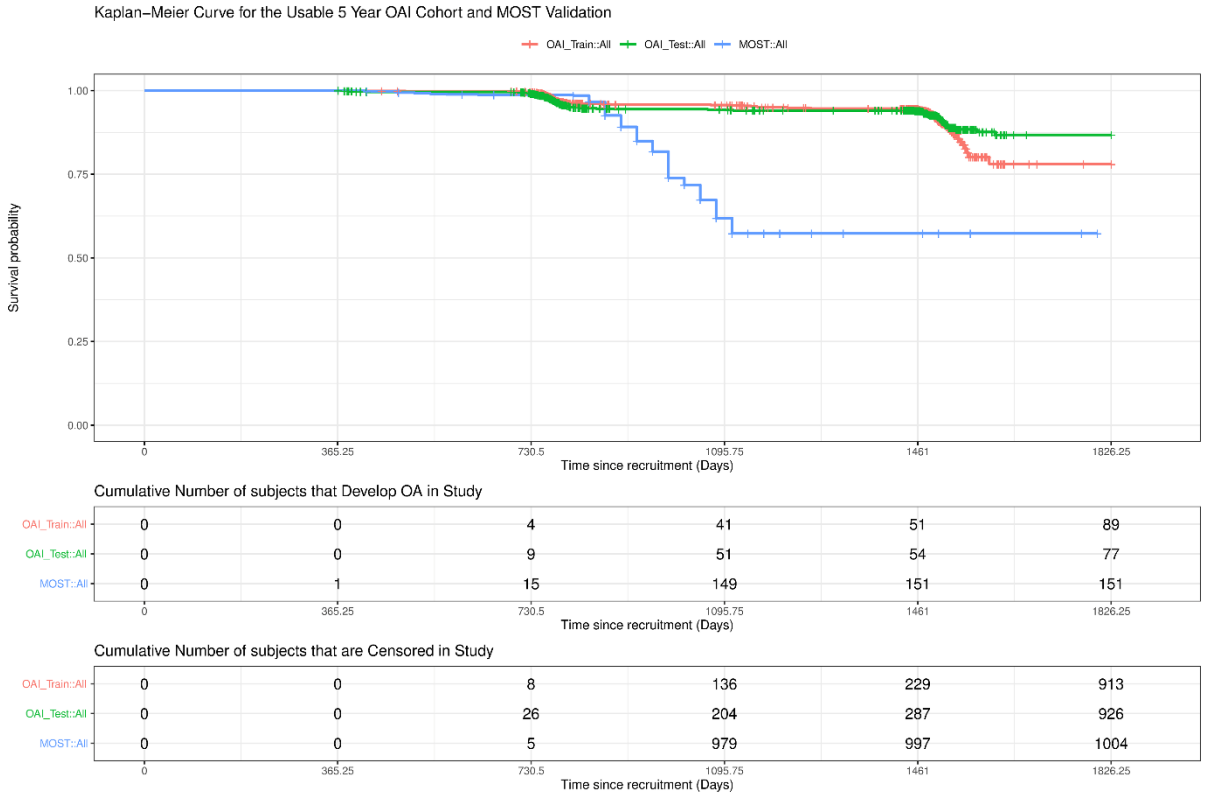


Figure 5-5: Observational KM curve stratified by sample. The red depicts the OAI training sample and the green shows the OAI test sample. The blue curve is the MOST validation data. The tables below illustrate the way in which the data is split between the samples.

To ensure the modelling of the variables is appropriate for the assumptions made about proportional hazards, testing is carried out. The results from these investigations show that all of the covariates, along with the model as a whole, follow the proportional hazards assumption.

The next step in the analysis is to see if there are groups within the cohort, displaying different risk profiles. For example, to determine if there is a high and low risk group, and to establish what the criteria is for inclusion in each group. To stratify the group into cohorts the first step is to establish a cut point that gives the biggest separation in the subjects.

Figure 5-6 shows the stratification curves on the OAI training data for the raw data and the predictions produced on the MOST validation. The last event recorded in cohort 2 on the training data within the 5-year span is at day 1642. The stratification curves produced are well separated with no crossover on the confidence intervals, which indicates that on unseen data the well-separated groups hold true.

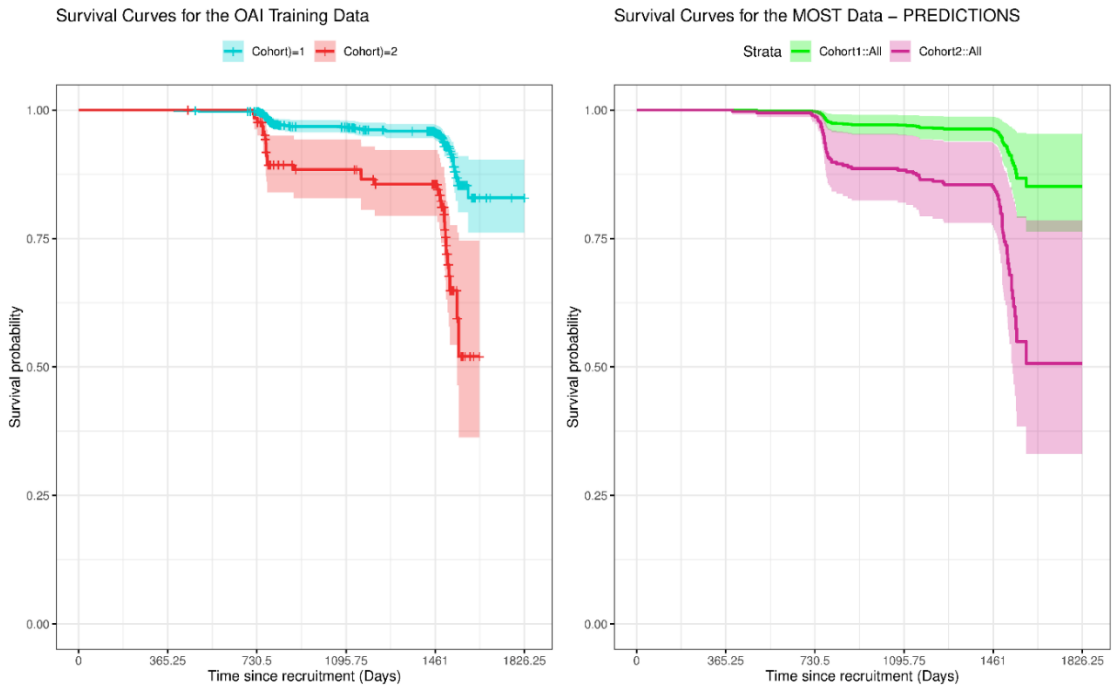


Figure 5-6: Stratification curves on the left showing OAI training data showing the high and low risk cohorts. Note the last event recorded in cohort 2 within the 5-year span is at day 1642. The stratification curves on the right are the validation data showing the high and low risk cohorts fitted to the models developed using the OAI data.

For the model to have clinical value, the findings of the two risk cohorts need to be translated into human terms. For example, how the features influence that individual in relation to which risk group they will belong. The proportions are shown in Figure 5-7. The proportion plots are useful as they can be used easier to profile the groups in each cohort. For example, in Cohort 2 the majority of the subjects are female, all with a BMI over 25, and the majority have had previous knee injuries and a history of falling. The majority of the people in Cohort 2 have no family history of knee problems, which could mean that those who were aware of the issues with their family history of OA had already made changes to their behaviours and this helped with prevention or delay in developing KOA. When calculated, the AUC for the survival analysis for the OAI test set is 0.74 (0.7325 – 0.7439) and that of the MOST data is 0.72 (0.7190 – 0.7373).

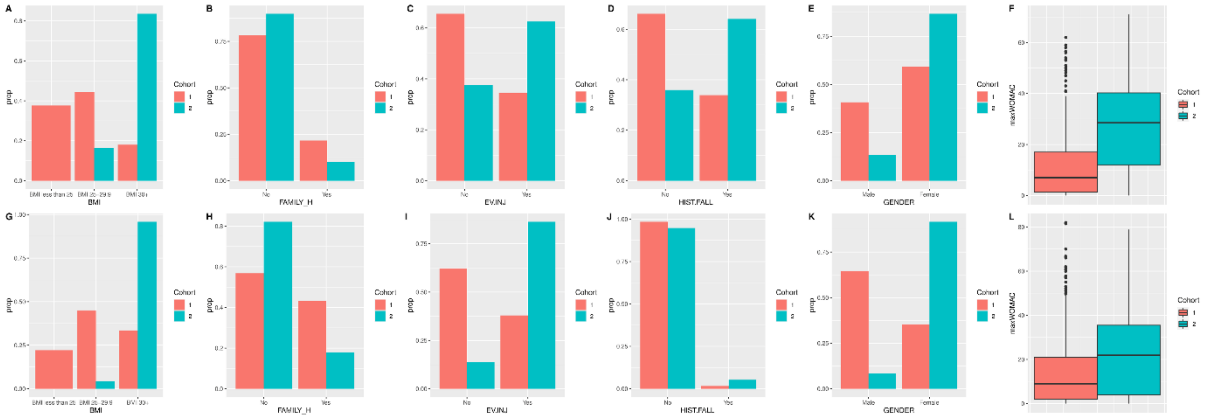


Figure 5-7: The cohort profiles per variable for the different strata. The red bars show cohort 1 and green show cohort 2. This representation of the profiles is the proportion of the group in each data category per cohort, graphs A-F are for the training set, and G-L are for the validation data.

For the prognostic and diagnostic OA prediction models to be useful in a clinical setting they need to be user friendly, and implemented in a digital format such as a web based app. Both the diagnostic and prognostic apps are shown in Figure 5-8 and Figure 5-9 respectively.

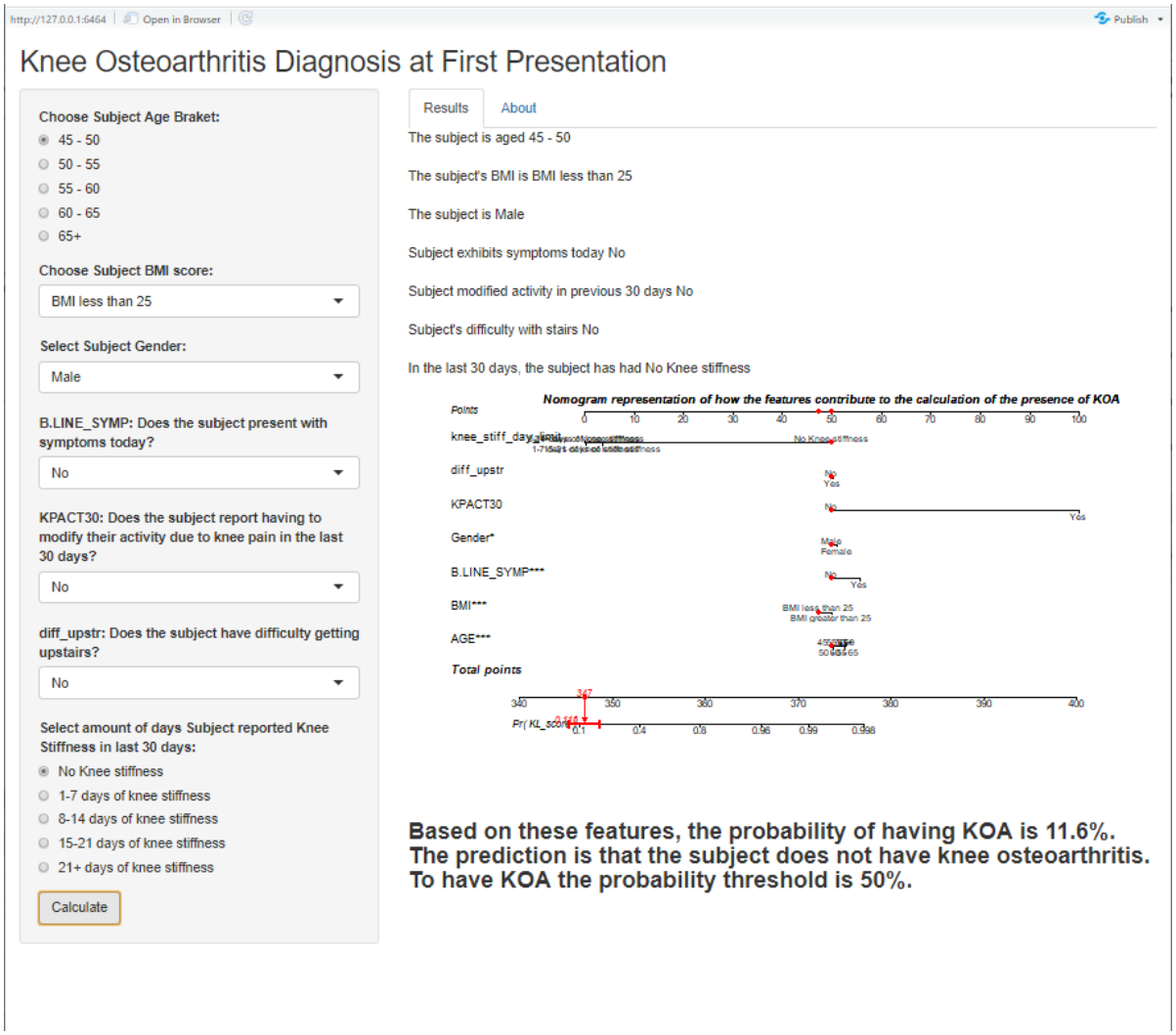


Figure 5-8: Diagnostic web app interface built in R Shiny. The app has multiple-choice options to allow the user to input variables that provide a probability of the participant having KOA based on the provided symptoms.

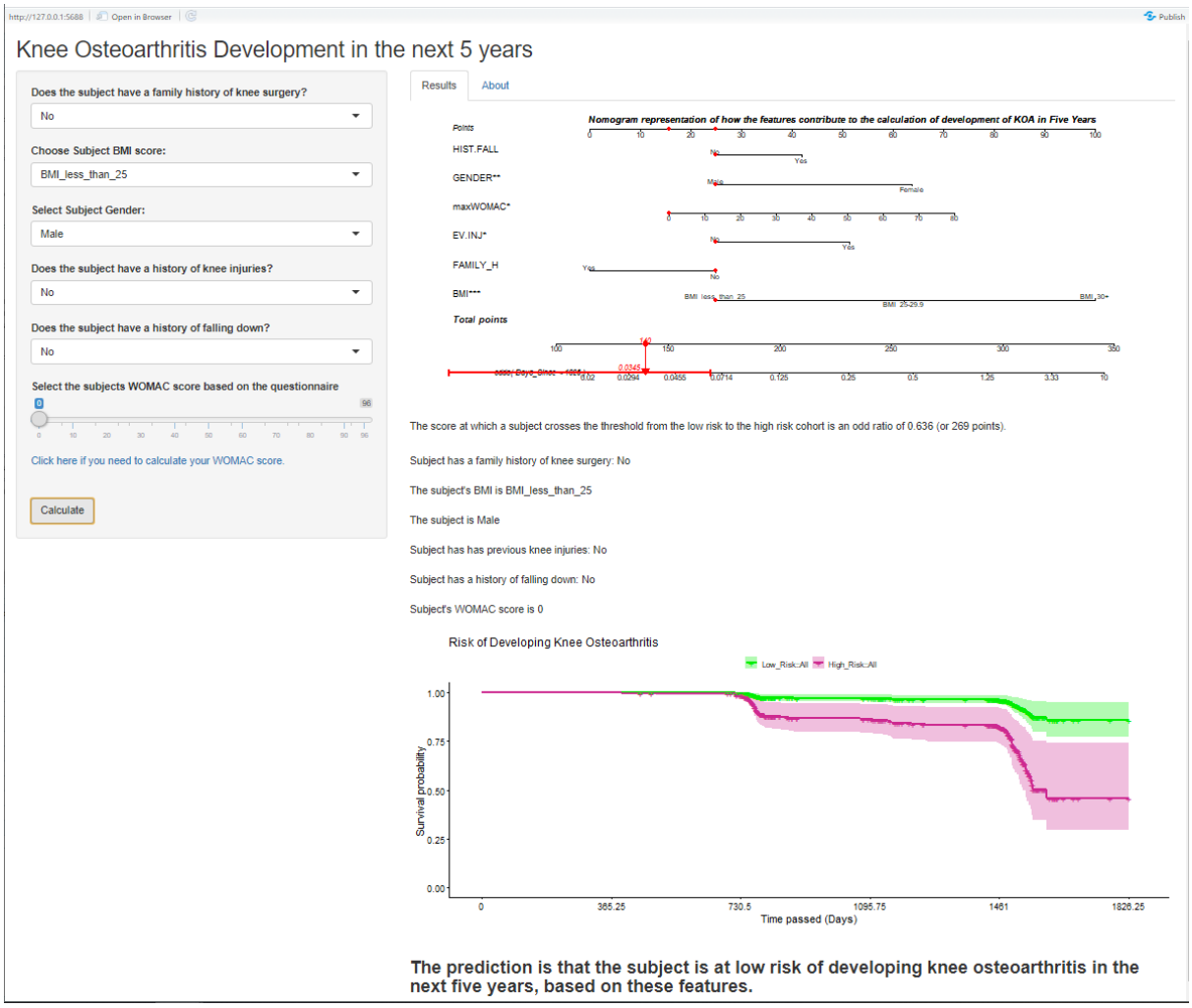


Figure 5-9: Prognostic app interface built in R Shiny. The app has multiple-choice options that relate to symptoms linked to the progression from a disease free state to KOA onset. The app also includes the option to link to the WOMAC questionnaire, should the participant not know their score.

5.5. Discussion

Both the prognostic and diagnostic models stood up to external validation with the MOST data. The OActive external validation on the diagnostic model did not produce a meaningful result due to the amount of missing variables in the data.

The success with validating on an external dataset, such as MOST, helps to ensure that the model has not overfit to the training data. By having data in the OAI study come from different centres and using validation data that was collected from different sources, we have greatly reduced any chance of the model overfitting to the noise in the data. The approach of using observational features to determine the probability of presence and likelihood of onset of KOA has not been considered before in this way, as this approach uses a selection of variables from different domains about an individual.

The OActive data is collected across three centres where recruitment criteria differed greatly. This had an implication on the number of cases present in each centre. In the cases for the centre UNIC, 62 subjects should be predicted to have KOA, and the model has predicted no KOA. Similarly, HULAFE had five cases that should be KOA and the model has predicted no KOA. Those subjects who were misclassified are likely to not suffer with symptoms, based on the small amount of data available, and therefore would not have symptomatic KOA, only radiographic KOA.

The use of decision support tools in clinical situations has filtered into many different areas. A 2012 review showed that the implementation of decision support systems were greatly effective at improving the processing for which they were create [161]. Decision support tools are used frequently when related to cancer. For example, Adjuvant! is a computer program developed in 2001 to allow health professionals and patients to make informed decisions about treatment [162]. This application was for use once a patient had a cancer diagnosis but helped to provide useful insight for the patient into the steps involved in decision making related to their care. A similar application was produced by the International Ovarian Tumour Analysis. This app is for clinician use, helping to determine if a tumour is benign or malignant. There have been two versions of this app with different rules created. One version was created in 2008 and uses six predictors in the model [163]. The later version of the app, from 2016, calculates a risk of malignancy in tumours, using the original model as a base that was modified [164]. Apps such as those, including the one developed for KOA in this work, (see Figure 5-8 and Figure 5-9) can be used in GP practices when a subject has symptoms or as part of screening to help educate the subject about their risk.

When looking at the area of knee osteoarthritis, survival modelling has predominantly focused on progression from an arthritic state to joint replacement. One example of this examined the importance of cartilage defects in older adults in relation to progression to knee replacement [115]. A similar study investigated the incorporation of radiographs when predicting the likelihood of total knee replacement within 9 years and the final Kellgren-Lawrence grade [116]. Some studies focus on the likelihood of developing KOA following certain treatment courses. For example, one such study examined the risk of requiring knee replacement surgery following treatment with intra-articular corticosteroid injections [117]. A similar study found that the use of intra-articular corticosteroid injection increases the risk of KOA progression [118]. Joint space narrowing was also

studied as an outcome in survival modelling in patients with known symptomatic OA [119] showing that once radiographic changes were visible then the risk of progression in OA was significant. This is where the prognostic modelling approach in this chapter differs, as it takes the subjects with no initial KOA and tries to identify those at a higher risk of developing the disease in a five-year follow-up window. This offers the chance to target healthy, at risk, individuals before the onset of KOA, and delay the onset of disease and also having the potential to reduce costs to the healthcare providers as treatment interventions may not be required as frequently as a result of educating the individual about their risks.

When considering the OAI models with OActive validation, we have developed two models: one for diagnosis of radiographic KOA and the other for progression from a disease free state to radiographic KOA in a five-year time span. The models were developed specifically for radiological KOA, assigned by a clinician with a KL grade. When modelling we have removed the potential for repeated measures as our model considers only the worst case for symptoms and outcomes. We also removed the risk of bias by imputation through having a complete cases analysis. The reasoning for this is discussed earlier in Chapter 2, section 2.5. The progression model defines a high and low risk cohort for developing KOA over a five-year window. While we considered KOA as a binary variable, future modelling could consider more granular changes in KL grade. Finally, when looking at the model performance, the 2016 model from [165] uses ANN and LogR, and on the externally validated data their ANN AUC is the same as our LogR AUC, which outperforms the AUC of their externally validated LogR model [165]. Also worth noting is that both their model and our models' internal AUC scores were roughly equivalent for our model and the PLOS One model [165].

One approach to consider for further expanding the model would be to use UK-based data and validate the model built on the OAI data [61]. This would be a useful way to check that models built with different demographics do translate to different populations. This would also externally validate the model for the UK population.

Chapter 6: The influence of gender when considering diagnostic modelling of knee osteoarthritis.

6.1. Introduction

About 50% of the global population is female yet in many areas, including medicine; women are often considered as inferior versions of men [166]. This idea dates back as early as Greek times, when Aristotle made reference to a female being a mutilated male [166], and in many ways this in western medical culture this view had remained. Published studies from the 1970's and 1980's on the use of aspirin to prevent heart attacks are examples of clinical trials that exclusively recruited men, where the conclusions did not hold true for women [167]–[169]. Despite many publications since that have urged the inclusion of females in clinical trials and for analysis to be conducted on populations by sex, this is still not always the case [170], [171]. Prior to 1993, when the Women's Health Equity Act was passed which gave women the opportunity to participate in medical studies, there was no research into conditions that are prevalent to women's health [172], [173].

In medical diagnosis there is a known phenomenon wherein there is a discrepancy in the way male and female pain is treated [174]. There are many cases where females are dismissed or told that they are imagining the problem, because not all illness and disease manifest identically in males and females [175]. Where diseases have physical symptoms many also have psychogenic symptoms, and many of these symptoms such as depression and anxiety, are conditions themselves [175]. Given that females are more likely to suffer from these conditions it is not shocking that in America they are twice as likely as males to have a diagnosis of depression or anxiety [176]. The societal idea that males are stoic and can handle situations whilst females are at the mercy of their emotions may have fed into the over diagnosis of depression in females and an under diagnosis in males [177]. Therefore, prevalence rates may have been partially influenced by this idea. Studies from the 1990's have shown that of the females diagnosed with depression as many as 30-50% were misdiagnosed [178].

Many females also experience a delay in receiving a diagnosis when compared to males. On average, it takes longer to diagnose females with the same condition as a male for many reasons including the disease manifesting differently [179]. One example is that it takes an average of 12 months for males to be diagnosed with Crohn's disease and 20 months for females [175].

Differences by sex can be and are significant. Researchers have found differences due to gender in every tissue and organ system in the human body [180], as well as differences

in the way diseases affect people of different genders [181]. There are differences in the mechanical workings of the heart that are due to gender [182]. This may help to explain why heart attacks in women are different from those in men [183]. 8% of the population have autoimmune diseases [184]but about 80% of those affected are female [185]. Researchers have a theory that the immune system in women is 'heightened' to protect a developing foetus [186], meaning that at times it can overreact and attack the body [187]. It may be because of this that women typically have higher antibody responses to vaccines and suffer more frequent and adverse reactions than males [188]. A 2014 study suggested a need for developing gender specific versions of the influenza vaccine [188]. Gender-neutral doses for the majority of medication including anaesthetic and chemotherapy continues to put women at risk of overdose [169], [189].

There is a known issue with treating every subject, male or female, as a generic subject. Examples of this type of gender bias are not only prevalent in medicine but reach out to other areas, with things such as how PPE is made and the average office temperature, with females 'suffering' in both of these scenarios. In many cases, the models that are used in medical settings have been trained in unevenly represented populations, with the demographics of the data leaning toward male dominance.

It is known that the prevalence, incidence and severity of osteoarthritis differ in women and men [190], with women more likely to suffer than men [191]. Women aged 50-60 may be 3.5 times more likely to develop osteoarthritis than men in the same group [192]. In addition to this, women are 40% more likely to develop KOA than men [193], and are also more likely to experience more severe knee osteoarthritis [191], [194]. Typically, when women present to a clinician with OA they are in the advanced stages of disease and are suffering from more debilitating pain [195]. It has been noted for many years that there are significantly higher number of women with symptomatic disease than their male counterparts [196]. However knowing that there is a higher incidence in women than men has not triggered a study to investigate the cause for the difference in presentation and development of the disease [195].

Despite no definitive reason for these differences, there has been some speculation around the causes [197]. Hanna et al. [198] suggests that the differences in male and female anatomy may contribute to the differences in disease manifestation. Bone density and size differences may contribute to the way in which men and women develop KOA. It has been found that at baseline, women are more likely to have higher numbers of

cartilage defects and then go on to suffer increased loss of cartilage than men, resulting in women being more likely to have an increased risk of developing defects of the cartilage; a risk factor and sign of KOA [199].

Kinematic difference between males and females may also contribute to the development in women [195], [200]–[204]. Female athletes load their joints with additional stress which can increase the risk and likelihood of developing KOA [195]. Women's knees have muskoskeletal differences which can also affect loading and wear [205], [206]. In addition to this, the way that injuries are managed and followed up can also contribute to the development of KOA [207]. Damage to the anterior cruciate ligament (ACL) and subsequent injuries that follow from this can lead to OA in both men and women [208], [209]. However, it is also known that women are more likely to damage the ACL [195]. Based on a study by Chu at al deploying early arthritis interventions following ACL damage may have the potential to positively impact the health of both men and women and their knees, further reducing disease burden on women who are at a greater risk [207].

Among the other factors that contribute to the onset of KOA, hormone levels have a role in disease development. A woman's susceptibility to OA may be related, in part, to hormone levels [210]. After the menopause, women are at an increased risk of developing OA. This has been linked to a drop in oestrogen, as there are oestrogen receptors present in cartilage [211].

The work in this chapter build on the existing models from Chapter 3 and the models that have been successfully validated by the MOST data, described in Chapter 5. The aim of this adaptation in the modelling process will consider the role that gender plays in the diagnosis of subjects with KOA by trying to expand on the existing knowledge relating to the disease diagnosis in males and females. By considering the model without gender as a variable we can establish whether a global model, with and without gender, is suitable for the whole population, or if there is a benefit for considering male and female subjects separately, with gender specific variables available for inclusion into the model.

Chapter aims

- Establish whether the global model performs equally as well for males and females when considered separately, as well as when gender is excluded from the model.
- Determine whether the introduction of gender specific variables alters the model for both the male and female cohorts.

6.2. Study Design

When looking at how best to implement the models to consider gender three approaches became apparent.

The first approach is to split the original data pool by gender. Take the original pool of variables and build a model for each gender. Use the respective data test sets to assess the performance of the model. Finally, develop a 'comparative' model with all subjects, not including gender as a variable, to assess the global model performance. The model performance between the global model and the separated models can be compared to determine whether gender is influential when predicting the presence of KOA.

The second approach starts the same as the first, by splitting the data pool by gender. Then, for the male and female cohorts add the gender specific variables to the respective set. On each of the expanded variable sets, do feature selection to determine the most influential variables. For each gender then build a model using the selected features. Assess the individual model performances and compare these to the original model. This will also highlight if the inclusion of gender specific variables is beneficial when considering males and females separately for predicting the presence or risk of onset of KOA.

The final approach is to consider only the variables that were determined to be consistently significant in the modelling. To select those variables, a comparison across the male and female features highlighted which variables were present in both male and female models. This limited set of variables was then used to build three models; one for all subjects, one for males and one for females. For each model, the performance was assessed and a comparison made to determine if there is any value in having separate criteria for diagnosis of KOA in males and females.

6.3. Gender specific factors used in the analysis

In both the diagnostic and prognostic models, Gender has been a variable selected due to its importance to the model. In the prognostic model, the majority of the high-risk cohort were female. If the two genders were considered separately with the addition of gender specific information, then the final model and the subsequent variables calculated to be significant for each gender may be different from those of the global model.

Using the OAI data, Chapter 2, several gender specific variables were identified and extracted. The original cohort was then split by gender and the additional variables added to each cohort respectively.

The female specific variables cover if the subject has ever had a hysterectomy, if they have had ovaries removed, if they have ever been pregnant or if they have ever had hormonal treatment for either menopause or bone density issues. The male variables consider the use of testosterone in the six months prior to data collection, or if the subject experienced prostate cancer. The variables added to the models for male and female subjects are described in Table 6-1.

Table 6-1: A table describing the gender specific variables included in the models, the data label, and the definition of each variable. Each variable mentioned in the table has a binary outcome, either yes or no.

Gender	**Variable**	**Definition**
Male	P02CNC13	Has the subject got or previously had prostate cancer?
	V00TEST	Has the subject used male hormones in the previous 6 months (prior to initial visit)?
Female	P02HYS	Has the subject had a hysterectomy?
	P01OVREM	Has the subject ever had an ovary removed?
	P01PREGEV	Has the subject ever been pregnant?
	V00ESTR	Has the subject ever used a combination of oestrogen and testosterone as either treatment for menopause or to increase bone density?

The OAI data contained other variables that could be of interest when considering their impact in relation to gender and the onset of KOA. These additional variables include whether the subject had ever had breast cancer and then female specific variables that were not included are whether the subject had cervical cancer, whether the subject had uterine cancer and how many ovaries were removed.

The first variable that could have been of interest was whether the subject had ever been diagnosed with breast cancer. Breast cancer is not a gender specific cancer as it can affect anyone in the population. However, the limited information present in the data meant that this variable was unsuitable for inclusion in the analysis. There were 19 subjects out of the total 2707 who had received a breast cancer diagnosis, and all of these were female. The only other state present for the variable was 'NA' which cannot be categorised as either yes or no. Therefore this variable was rendered unusable for this analysis.

The next variable that proved unusable was whether the subject had a diagnosis of cervical cancer. Cervical cancer is a type of cancer that can only affect people with a cervix and

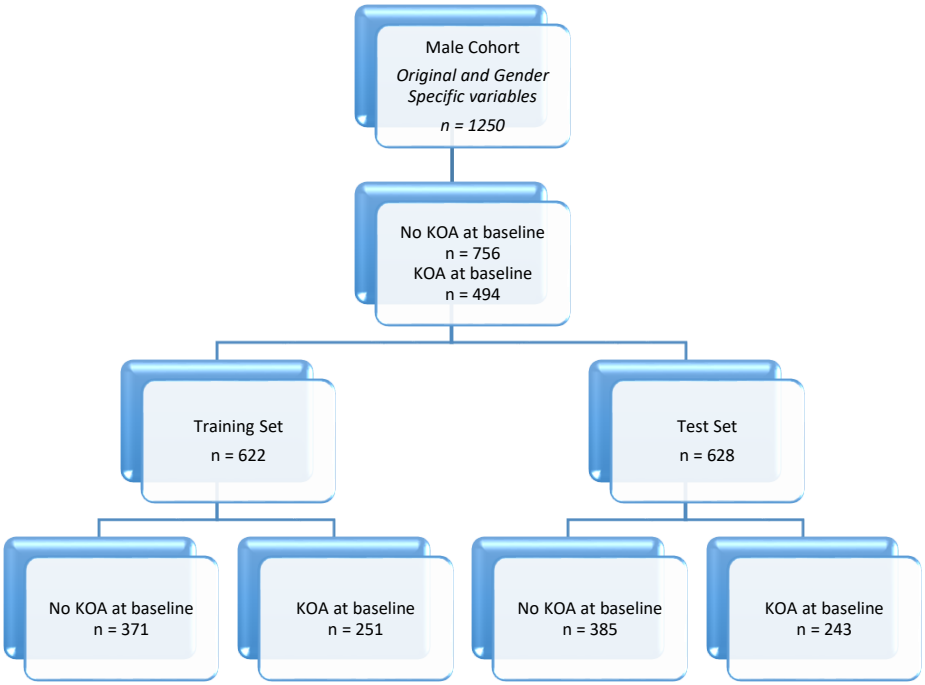
therefore, if included, would be considered in the female modelling only. The problem with this variable in the OAI data is that for the cohort with the original variables that are used, all subjects have 'NA' values, rendering this variable unusable.

There was a variable that considered a diagnosis of 'colon, uterine or other' cancer. The problem with this is that these options are grouped, leaving no scope to differentiate between the specific cancer type, meaning that the presence of 'yes' could potentially be such for uterine cancer, a female specific disease, or colon cancer, which could affect anyone. For this reason, this variable was not considered for the analysis.

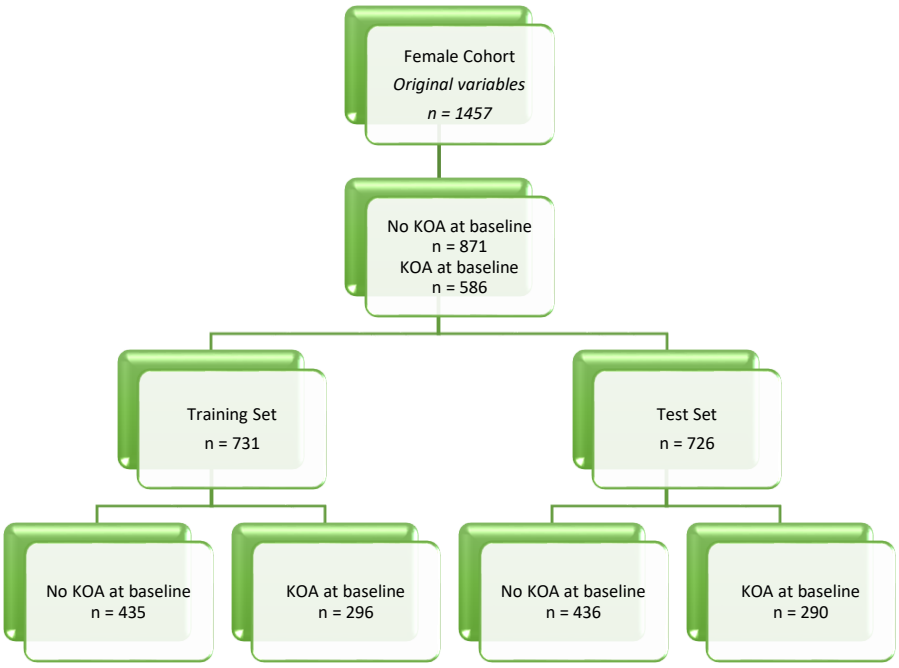
The final variable that could be of interest in the gender specific analysis is how many ovaries have been removed. As this chapter details the preliminary investigation, its purpose was to identify if disease prediction can be improved by considering gender specific factors. For this reason the binary version of this variable 'Have you had ovaries removed?' was used instead of 'How many ovaries have you had removed?'. The latter option could provide an initial step for future investigation into what, if any, difference to disease presence occurs as a result of removing one, both or no ovaries. A subsequent step could be to explore the impact of the type of hysterectomy a person has in relation to KOA, as there are three types, with only one resulting in the removal of the ovaries.

The data in this chapter is specifically grouped by gender for the analysis, resulting in two cohorts: male and female. In each of these cohorts there are training and test splits to allow the model performance to be assessed. The male cohort consists of 1250 subjects, with the prevalence of disease at 0.4. The female cohort varies by study approach, due to some subject having missing values for the variables of interest. The original variable approach gives a female cohort size of 1457, and the inclusion of gender specific variables reduces the cohort to 1442 subjects. The way the in which the data is split for modelling and testing is detailed in Figure 6-1.

A


Male Cohort
Original and Gender Specific variables
n = 1250
No KOA at baseline
n = 756
KOA at baseline
n = 494
Training Set
n = 622
Test Set
n = 628
No KOA at baseline
n = 371
KOA at baseline
n = 251
No KOA at baseline
n = 385
KOA at baseline
n = 243

B


Female Cohort
Original variables
n = 1457
No KOA at baseline
n = 871
KOA at baseline
n = 586
Training Set
n = 731
Test Set
n = 726
No KOA at baseline
n = 435
KOA at baseline
n = 296
No KOA at baseline
n = 436
KOA at baseline
n = 290

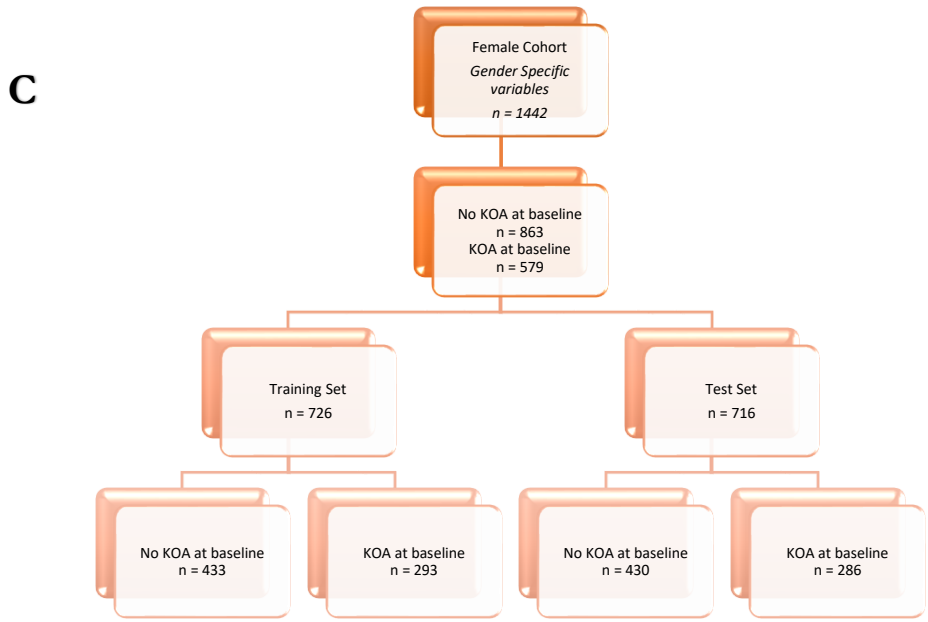


Figure 6-1: Data breakdown of the cohorts detailing the prevalence of KOA in each population and the training and test set sizes. (A) shows the data breakdown for the male cohort for both the original and gender specific variable approaches, (B) shows the female cohort breakdown for the original variable approach and (C) shows the female cohort for the gender specific variable approach.

6.4. Diagnostic Modelling Results at Baseline

The data in Table 6-2 show a results summary of the models used and their performance values to assess what is the most beneficial way to incorporate gender into diagnostic modelling. Rows 1, 2 and 7 show the results for the models considering all variables. The three AUC values are all comparable with overlapping confidence intervals. The same can be said for the male and female specific models. They are all consistent in having high specificity values. The models therefore would be good at identifying cases where KOA is not present. In each case where there are male and female specific models, the female model outperforms that for the male. This is a good indication that there is a need for sex-disaggregated modelling within medical applications.

Table 6-2: A table of performance metrics for the different variable sets used for the analysis, giving the area under the curve (AUROC), sensitivity, specificity, and positive predictive value (PPV).

		AUROC (CI)	Sensitivity	Specificity	PPV
1	All subjects with gender	0.7481 (0.7214 – 0.7748)	0.5197	0.8477	0.7734
2	All subjects without gender	0.7483 (0.7217 – 0.775)	0.5084	0.8502	0.7724

3	Female model original data	0.7595 (0.7237 – 0.7953)	0.4966	0.8899	0.8185
4	Male model original data	0.7272 (0.6867 – 0.7677)	0.4979	0.8130	0.7270
5	Female model original + gender variables	0.7669 (0.7313 – 0.8026)	0.4895	0.8977	0.8271
6	Male model original + gender variables	0.7248 (0.6842 – 0.7654)	0.4815	0.8208	0.7287
7	All Subjects Significant Variables	0.7516 (0.7252 – 0.7779)	0.5197	0.8404	0.7651
8	Male Significant Variables	0.7285 (0.6882 – 0.7688)	0.4979	0.8130	0.7270
9	Female Significant Variables	0.7648 (0.7294 – 0.8001)	0.4931	0.8945	0.8238

Looking at the results in Table 6-2 it is clear that there is a small amount of variation in model performance. Consistently, the best performing models are those trained and tested on only the female cohort. It is understood that there is likely a link between gender and risk of onset of KOA. Although the differences between the male and female only models are not significant, the small differences that are present highlight a potential need for a dedicated analysis with a larger sample size and more variables to consider if gender does play a significant role in the likelihood to develop KOA.

The variables considered in the model from row 1 in Table 6-2 is the same as the model described in Chapter 3. Rows 2-4 consider the variables Age, BMI, baseline symptoms, knee pain impacting activity in the previous 30 days, knee swelling, difficulty getting upstairs and limiting knee stiffness. The data was modelled with the whole cohort, results in row 2, and then split by gender, rows 3 and 4. The ROC curves for these models are shown in Figure 6-2.

Rows 5 and 6 detail the results for the curves shown in Figure 6-3. These models consider the original pool of variables plus the gender specific variables, and only those chosen by feature selection are included in the model.

The female model uses age, BMI, baseline symptoms, knee swelling, limiting knee stiffness, history of pregnancy and history of hysterectomy. Two of the four 'female' variables were chosen by feature selection to be included in the model. Although neither

of these were calculated to be significant in the model, they still contributed to the prediction.

The variables in the male model were age, BMI, baseline symptoms, knee swelling, limiting knee stiffness, difficulty getting upstairs and history of prostate cancer. One of the 'male' variables was chosen by feature selection and therefore included in the model for male subjects.

Figure 6-4 shows the ROC curve for rows 7-9. This model considered only the variables calculated to be significant in the previous analysis. The variables are age, BMI, baseline symptoms and knee swelling. Just considering these variables gave an improved performance over the original model, described in Chapter 3. However, the inclusion of the other variables can help provide more insight into the general wellbeing of the subject. In this analysis, the use of only the significant variables was only improved for females when considering the genders as cohorts.

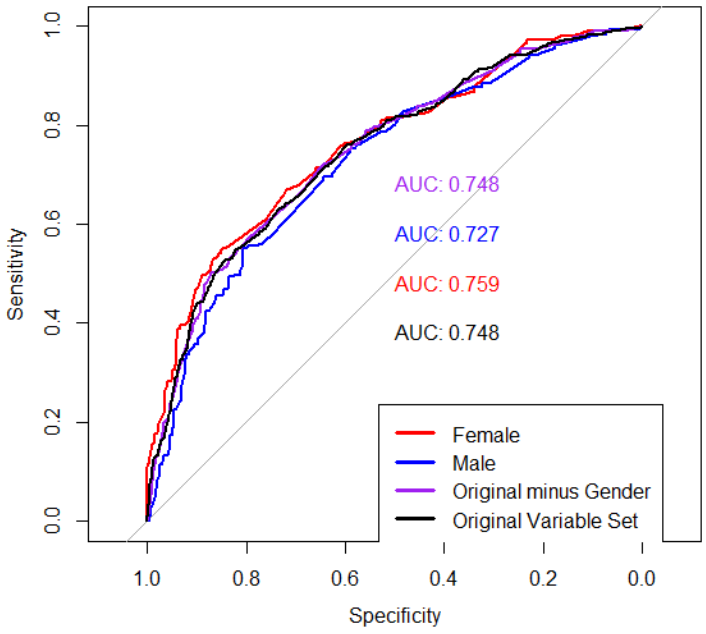


Figure 6-2: ROC curves for the original data split into male and female groups. The ROC curve for the whole cohort, with and without gender considered in the model is also present.

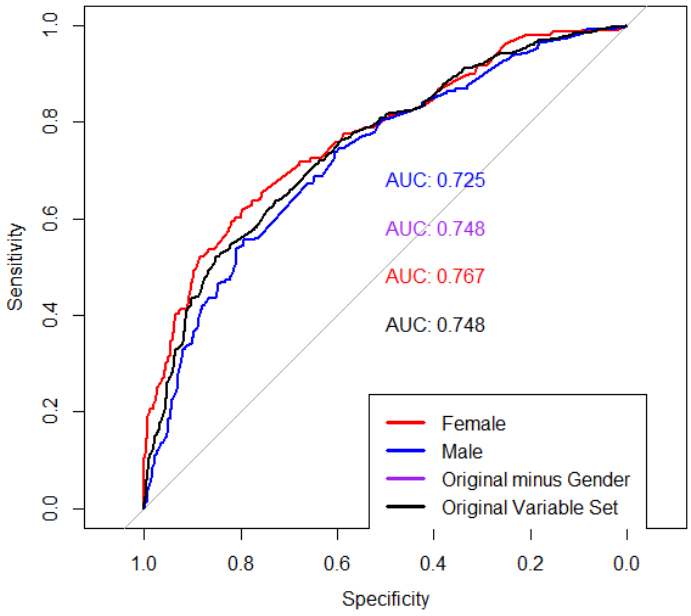


Figure 6-3: ROC curves for the original and the gender specific variables split into male and female groups. The ROC curve for the whole cohort, with gender considered in the model is also present as a baseline.

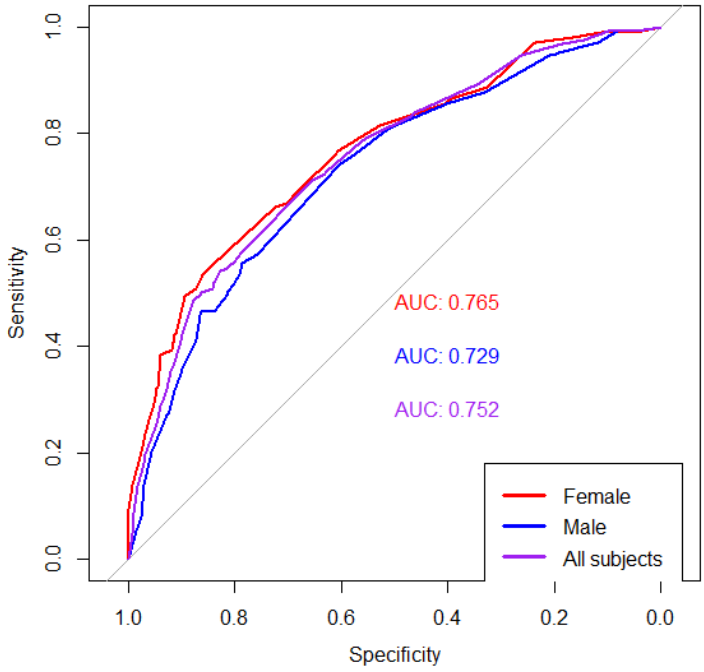


Figure 6-4: ROC curves for the analysis considering the significant variables only, split into male and female groups. The ROC curve for the whole cohort is also present as a baseline.

It is clear from the gender specific nomograms, female in Figure 6-5 and male in Figure 6-6, that the variables contribute to the likelihood of having KOA differently when considered solely for males or females compared to the whole sample.

Considering Figure 6-5 and Figure 6-6, it is clear that BMI has a greater influence on a males likelihood to develop KOA that for a woman. Similarly, knee swelling in males increases a male's chance of contributing to having KOA nearly double that for a female. Up to the age of 50, age contributes to female disease likelihood more than for males, but from age 55 years age contributes more to males, with 65 year old males having an equal age related contribution to the overall KOA likelihood as a 75 year old female.

Considering the model without gender, displayed in Figure 6-7, and with gender, displayed in Figure 6-8, there appears to be little-to-no difference between the variable contributions to likelihood for disease presence. This is supported by the AUC for these models, all being 0.748 to three decimal places.

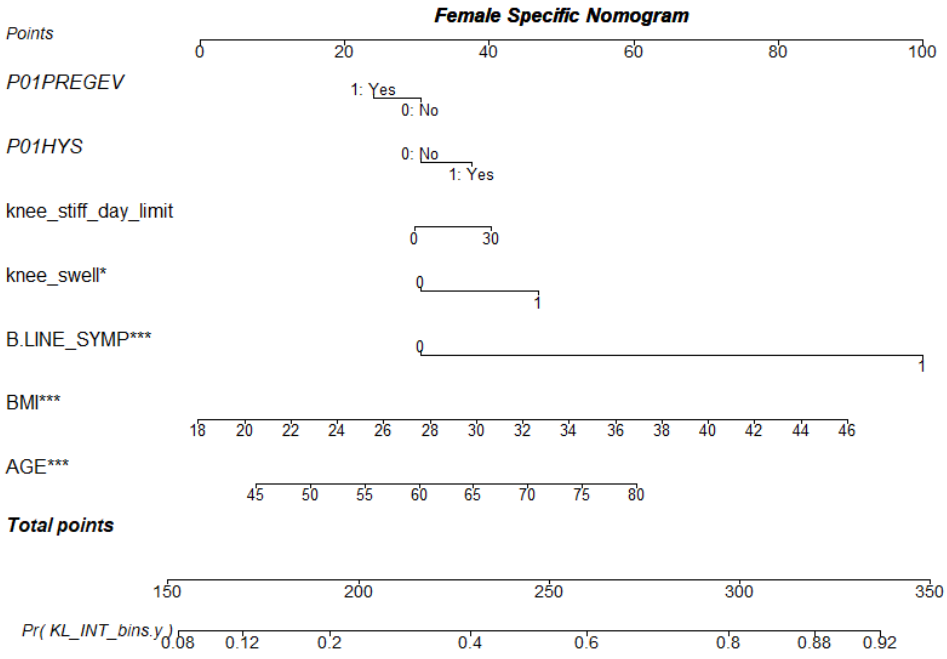


Figure 6-5: A nomogram showing the Female specific model for diagnosing KOA at first presentation.

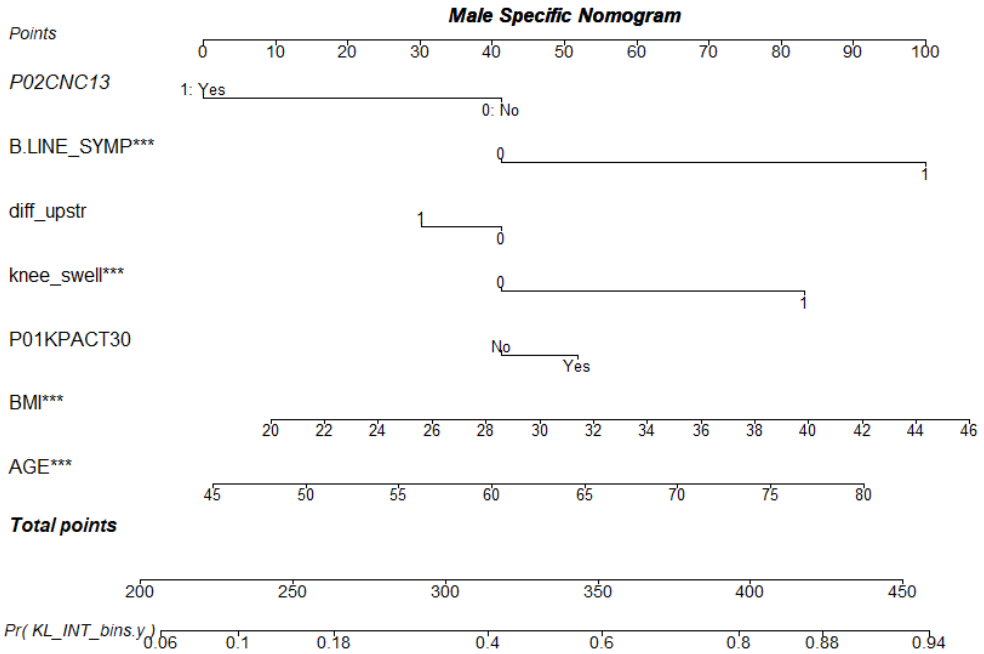


Figure 6-6: A nomogram showing the model calculated using the male specific variables to diagnose KOA at first presentation.

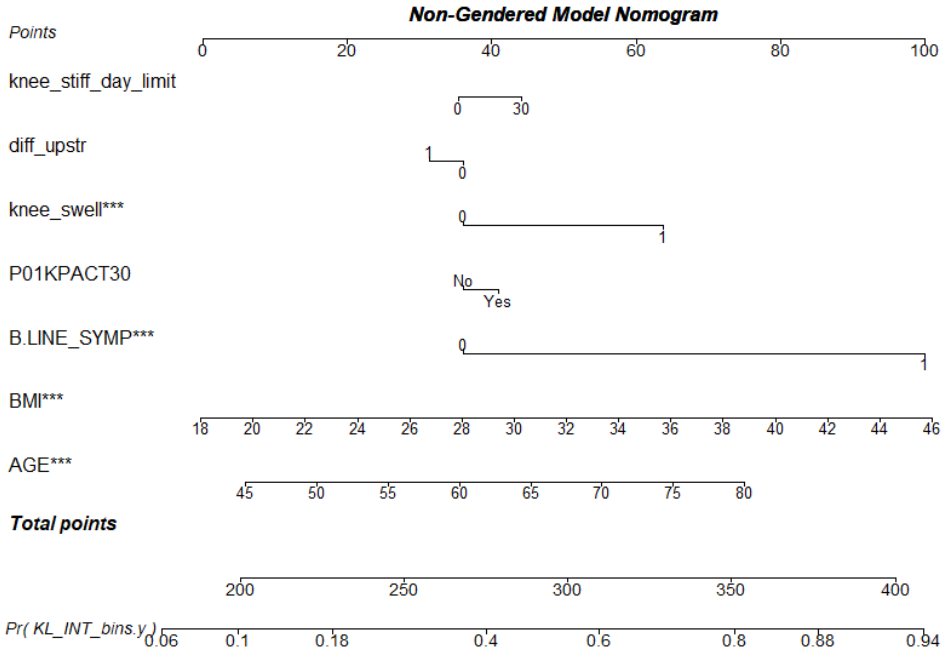


Figure 6-7: The diagnosis model for KOA calculated after removing the gender variable.

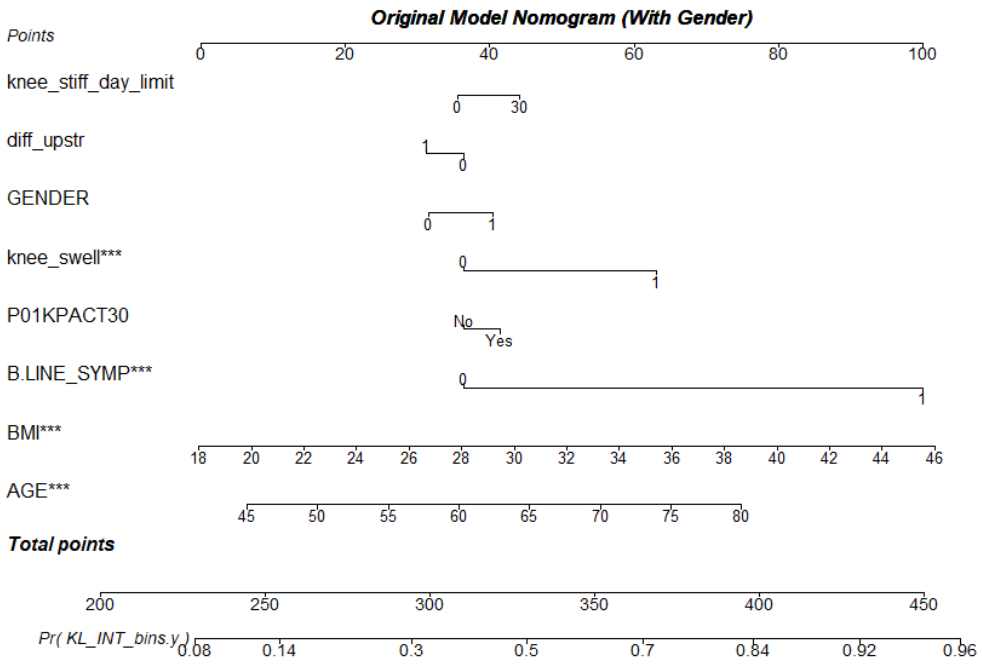


Figure 6-8: A nomogram displaying the variable contribution of the original diagnosis model, described in Chapter 3, to the chance of having KOA at first presentation.

6.5. Discussion

The work in this chapter only highlights the current understanding that more research into the effect of hormones on the development of knee osteoarthritis is required [195]. There are already established disparities in knee osteoarthritis due to gender and the differences this can cause both biologically and behaviourally [212]. Despite a growing pool of evidence that supports this finding there is a lack of translation from this into practical implementations. This is leading to an unconscious bias from healthcare professionals when diagnosing and identifying those at risk from KOA.

As far back as the 1980's there have been more women affected by symptomatic KOA but this has not influenced how disease in men and women is considered. A risk factor for KOA is being female, but when roughly half of the global population is female that offers no comfort or insight into understanding the condition.

The studies carried out into the effect of gender on the overall benefits of surgical intervention show varying results, with no clear difference between males and females [213]–[215]. There is, however, evidence to show that after total knee arthroplasty female patients are significantly less likely to be satisfied with the pain relief they experience [213].

The analysis in this chapter shows that there is scope for further investigation into the differences from a diagnostic point when considering the way KOA affects females and males. Although there were slight differences in the performance of the models when considering male and female separately, the overlap of the confidence intervals suggests that there is no real significance in the differences between the models. However, the additional information leveraged by using gender specific variables also indicates the importance of interpretability as it may have helped to find underlying reasons for predictions instead of picking up surrogate factors that perform comparably, but without the added explanation provided by the specific variables.

Considering the female specific factors in the model, history of hysterectomy and history of pregnancy highlight a potential link to hormone involvement, specifically oestrogen. During pregnancy, the level of oestrogen in the body increases in comparison to the relative concentration in a person who is not pregnant. After a hysterectomy the level of oestrogen in the body falls, a pattern that is mirrored in women who have gone through menopause. The studies focusing on whether oestrogen therapy has a protective effect for development of OA have conflicting findings, with some suggesting protective effects, and others suggesting the likelihood to increase risk to develop OA [216]–[218]. One study identified that hormonal and reproductive factors have an effect of risk of knee replacement [217], suggesting that consideration into these features should be accounted for when determining if someone is at risk of developing KOA.

Even though the models containing gender specific variables do not outperform the generic model. The factors in the gender specific models are more specialised, so therefore could arguably describe the difference in performance. Although the gender specific models are not significantly different from the others, they are more specific, which in a screening scenario could be argued is more useful to both the clinician and the subject. One of the key ideas throughout this thesis is that of interpretability, and this additional information adds a layer of interpretability to the model that can help the subject better understand what is happening to them in a more specific way. The added insight can be provided visually by using nomograms, in this case, can also justify the advantage for using gender specific models as part of screening and patient education programmes.

Until further research is carried out into the effect of hormones, both male and female, into the likelihood of developing KOA there is likely to be an unconscious bias. Studies

have shown that there are genetic differences between males and females in every tissue in the body and until the effect of this is understood there needs to be a focus on engineering a 'best-fit' model for groups of the population, without overfitting to the trend highlighted in the data available at the time.

Chapter 7: The application of multi-task learning to diagnostic models for knee osteoarthritis.

7.1. Introduction

Multi-task learning (MTL) looks to solve multiple tasks at once by exploiting and making use of commonalities and differences in the data. In the early days of the field, MTL was known as hints [219], [220]. There are several different types of MTL; assistant task, piling, transfer of knowledge and group online adaptive learning. In one form or another, all of the approaches look to optimise the use of the available data to maximise the machine-learning model performance. When compared to single task analysis, where datasets are trained separately, MTL often produces improved efficiency and predictive accuracy.

7.1.1. Types of multi-task learning

7.1.2.1. Assistant Task

This method is also known as task grouping and overlap. The way in which this method works is that information is selectively shared across the different tasks based on a measure of similarity. The underlying theory is that the similarity indicates relatedness in the data. Relatedness in the data can either be determined from existing knowledge or can be learned newly from the data [221], [222].

In some research, this method has produced experimental results that outperform other, standard methods. When an improvement is not produced, and the performance has worsened, the experiment conducted is said to have had 'negative transfer' [223].

For this particular branch of MTL, there are two different approaches for sharing to occur [224]. The first approach is sharing in an explicit manner. This approach assumes that the models share common structures or parameters [225]. The only issue with this approach is that it can sometimes cause the model to overfit. The second approach is sharing implicitly. This approach can reveal hidden relationships among learners [223]. There does not seem to be a method that is proven best in all circumstances.

7.1.2.2. Piling

Another approach to MTL, and the one used in this analysis, is piling. The piling technique looks to exploit unrelated tasks. With this method, the job of learning new tasks is carried out with the use of a group of tasks unrelated to the principal task. These unrelated tasks are sometimes referred to as auxiliary tasks. The learning of the principal and auxiliary tasks is better when the same data is used with both, as idiosyncrasies in the distribution of the data are removed, giving rise to cleaner and more informative data representations. When this method has been used on both real and dummy data the

results have indicated that exploiting unrelated tasks can also cause significant performance improvements over standard methods of learning [226].

This approach builds on findings by Argyriou, [227], but adds a regularisation term to penalise the product between the predictors of any two tasks originating from different groups. As a result, this type of MTL has the power to distinguish what features are important for each task, resulting in an improvement in the statistical performance [226]. The piling technique is merely an extended version of the assistant task method.

7.1.2.3. Transfer of Knowledge

Transfer of knowledge is related in part to the concept of knowledge transfer. The shared representation that is present in the transfer of knowledge is sequential, whereas the representations developed by MTL in its purest forms are typically concurrent. A well-known method for machine-learning, deep convolutional neural networks develops strong representations which are then useful to other algorithms that are learning related tasks [228]. One example is where a pre-trained model is used as a feature extractor to pre-process for another algorithm [229].

The transfer of knowledge approach addresses three main challenges faced when conducting MTL [230]. The first challenge is making the model learning process incorporate the task selection. The next challenge is to make the model learn the shared information at the pace of the system. The final challenge is to ensure that the model can be generalised to fit a wider group of MTL problems.

7.1.2.4. Group Online Adaptive Learning

When the data in the analysis is stationary, meaning that it does not change over time, typical MTL approaches work, but if the data is non-stationary, where there are changes to variances and means over time, then group online adaptive learning is used [231], [232]. In this way, it has been useful to share information if the learners operate in environments that change over time. This is because the learner would be able to benefit from the previous experience of another learner to be able to adapt rapidly to the new environment. One well-known use is predicting financial time series data trends based on what has happened previously, and the micro trends, like a small crash in the market before a major one, that also have an impact on the overall trend of the data [233].

7.1.2. Transfer Learning

Transfer learning has been developed to assist in the transferability of different data problems. Transfer learning can be defined as the ability of a system to be able to

recognise and apply previously acquired knowledge to new problems [234]. By applying existing knowledge, it is possible to solve problems quicker than with standard approaches and is often the case also that the solutions are better with the use of transfer learning. Like the assistant task branch of MTL, the premise for transfer learning is that the tasks used to gain the knowledge and the area where the knowledge will be applied must be in some way related because without a relationship there will be no improvements in performance.

Similarly to MTL, transfer learning has different types. The four categories and the use of each are dependent upon which aspect of knowledge will be transferred. *Instance-based transfer learning* works by assigning different weights to samples based on their perceived importance to the problem [235]. *Feature representation* uses numerical coding to represent structural information relating to the problem [236]. Transfer learning *techniques using parameters* share parameter-held information to add knowledge required in the problem [237]. The final type of transfer learning is known as *relational knowledge transfer*. The final type maps data from the original source to the final destination in order to improve the performance of the machine-learning method being used [238].

7.1.3. Why Multi-task Learning Works

MTL was proposed by Caruana in 1998 [225]. The initial premise was laid out such that two tasks, A and B, share a common hidden layer representation, F. MTL increases the size of the sample of data being used to train the model. When the model trains on task A the goal is to be able to train a model to learn a dataset without the noise typically present in that given data. As different tasks and their affiliated data have different noise patterns the use of MTL promotes the learning of a more general representation of noise [239]. A may be the only task to overfit but tasks A and B mutually assist the model to get better representations, F, by averaging the noise present in the model.

MTL is sometimes used as a way to highlight features of interest. Many noisy or high dimensional datasets pose problems when trying to determine the features that are relevant, but MTL can be used to highlight the ones of particular interest. MTL used in this way is known as attention focusing, and it is used to help provide evidence as to the most relevant or irrelevant features in the data.

Another approach that works for aiding in learning multiple tasks is eavesdropping. Some features are easy for some tasks to learn and problematic for others. The MTL allows the

model to 'listen' to the other tasks being learnt in the model. The easiest way for this to take place is by making use of 'hints' [220]. These hints are then used to directly train the model to predict which features are the most important.

One of the features that make the MTL method work is the allocation of weight. MTL assigns weight to the more favourable tasks. Doing this helps the model to generalise in the future to new tasks. This is because the new tasks are using the same the hypothesis space used for the original tasks, provided that they are also from the same data environment. The introduction of a bias reduces the risk of overfitting as well as the ability of the data to fit the random noise.

7.1.4. When MTL can help

In situations where the data does not have a large number of cases MTL can help to leverage the information contained in the dataset by using related data in the model to help with the learning process. This was an early motivation for the use of MTL, to alleviate data sparsity as a result of limited suitable data points [240].

In smaller sized datasets, there is always the risk that the model will overfit to the noise present in the training data, resulting in poor performance in the test set. By using more data MTL approaches allow the model to learn more general representations of the tasks, leading to more powerful models, better performance and a lower risk of the model overfitting to the data [241].

Missing values leading to incomplete data also pose a problem when training a model as this can cause the performance to suffer. However, the use of MTL in chemoinformatics has found that the differences in model performance when comparing complete data with a model trained on data containing missing information was very small [242]. This could potentially challenge the viewpoint of 'more data is better' from the perspective of collecting more data when similar compatible data exists and is available.

Appropriately leveraging the information contained in datasets, even with state-of-the-art methods can often reach a ceiling in performance. In order to exceed this maximum, one option is to utilise additional data. This can be done using MTL approaches. By enriching the pool of data, there is a chance that model performance will be improved.

7.1.5. Where MTL has been used

MTL has been used and shown improvements in many other domains. Its use in cancer drug research has been affiliated with improved prediction in the use of precision

oncology [243]. It also has many applications outside of biological science and drug discovery. One such example is spam filtering. The model would have key things such as language an e-mail has been written in and the language of the recipient's geographical location, but on its own that could exclude Spanish e-mails from people on holiday. So this, along with the language of the emails they usually receive and key phrases are used as filters to block unwanted and possible spam e-mails [244].

Another example of MTL being used is in web searching. MTL used with boosted trees has been used on web-search ranking data. This is especially useful when using data from different geographical locations. Using MTL here is especially helpful as data sizes can vary quite significantly due to cost. Learning tasks in this way has produced significant improvements in performance whilst retaining reliability [245].

In healthcare, a field with a vast amount of data, MTL has the potential to unlock more information in the data and boost model performances. This extra insight can highlight relations in the data that may have otherwise not been apparent. One example is by using a widespread collection of electronic health records there is an opportunity to utilise MTL approaches to develop more accurate personalised risk models [246].

Other application domains that benefit from the use of MTL include speech recognition, bioinformatics, computer vision and natural language processing [239], [240].

7.1.6. Scope for MTL use

The use of MTL in this thesis builds on the model from Chapter 3 with the aim of determining whether the use of MTL will have a positive impact on information transfer and therefore model performance. By utilising multiple data sources, the OAI and the MOST datasets, there is potential in the capability to improve model performance by solely using available data. This work takes a speculative look at the application of MTL to KOA diagnostic models.

In this way, for this preliminary work into the application of MTL to the KOA diagnostic model, the simplest MTL approach, piling, is used. By using the piling approach for MTL, it will be possible to establish if there is worth in the application of MTL for the purpose of enriching a data pool to improve the models performance. However, as the two datasets in use are very similar there is no underlying expectation that the MTL approach will produce a performance improvement. This is, in part, due to the data having similar strengths and weaknesses as both datasets were collected in a similar way despite the

studies beginning conducted in separate regions of America, beginning around 12 months apart.

Chapter aim

- Determine if it is possible to improve model performance on the OAI/MOST data by utilising an MTL approach

7.2. Specifics of the data used in chapter

When considering the application and use of MTL in the diagnostic modelling of KOA a suitable pool of data was required. For this chapter, both the OAI and MOST data were used to build and test models for disease diagnosis. Initially, the same cohort that was used in Chapter 3 was analysed, and this was followed up with an extended variable set.

As the analysis used the same initial cohort, a complete case approach was also used in this set of experiments. This involved removing any individual subject that had at least one missing value in the data.

The rationale for using the original cohort was to determine if an improvement to performance was likely caused by the additional use of MTL. This approach, along with the work in Chapter 3, provides a baseline to compare any performance improvement.

The original OAI dataset used in this analysis is summarised in Table 7-1. The data consists of 4,433 subjects with information on eight features, including the outcome, the presence of KOA. The original MOST dataset is also summarised in Table 7-1. This dataset is comprised of 2,006 subjects with the same features present in the OAI data.

Table 7-1: The summary of the MOST and OAI datasets.

	MOST (N=2006)	**OAI (N=4433)**	**Total (N=6439)**
AGE			
- Mean (SD)	62.336 (8.139)	61.081 (9.173)	**61.472 (8.882)**
- Median (Q1, Q3)	62.000 (55.000, 69.000)	61.000 (53.000, 69.000)	**61.000 (54.000, 69.000)**
- Min - Max	50.000 - 79.000	45.000 - 79.000	**45.000 - 79.000**
BMI			
- Mean (SD)	31.105 (6.338)	28.574 (4.821)	**29.362 (5.466)**
- Median (Q1, Q3)	30.260 (26.642, 34.515)	28.200 (25.000, 31.700)	**28.800 (25.400, 32.500)**
- Min - Max	16.720 - 71.910	16.900 - 48.700	**16.720 - 71.910**
B.line_symp	1567 (78.1%)	1256 (28.3%)	**2823 (43.8%)**

KPACT30	1839 (91.7%)	1188 (26.8%)	**3027 (47.0%)**
knee_stiff_day_limit			
- Mean (SD)	6.461 (10.949)	3.520 (7.961)	**4.436 (9.100)**
- Median (Q1, Q3)	0.000 (0.000, 7.000)	0.000 (0.000, 2.000)	**0.000 (0.000, 3.000)**
- Min - Max	0.000 - 30.000	0.000 - 30.000	**0.000 - 30.000**
Gender	1264 (63.0%)	2574 (58.1%)	**3838 (59.6%)**
diff_upstr	1870 (93.2%)	2454 (55.4%)	**4324 (67.2%)**
KL_score	**1214 (60.5%)**	**1982 (44.7%)**	**3196 (49.6%)**

In addition to the original cohort, an extended cohort was also considered. The extended cohort considers variables relating to the presence of other conditions a subject may have. These conditions, such as presence of diabetes, are not known to be linked with the presence of KOA but have the potential to offer additional scope for screening subjects if there is found to be a relationship with these conditions and the likelihood of KOA.

The extended datasets have information on 19 different features of interest, including the original set. From the OAI dataset there is data on 4,004 subjects, shown in Table 7-2, and the MOST has 2,427 subjects fitting these constraints, as shown in Table 7-2 also.

Table 7-2: The summary table for the extended variable sets for both OAI and MOST.

	MOST (N=2427)	**OAI (N=4004)**	**Total (N=6431)**
AGE			
- Mean (SD)	62.276 (8.047)	60.947 (9.116)	**61.449 (8.751)**
- Median (Q1, Q3)	62.000 (55.000, 69.000)	60.000 (53.000, 69.000)	**61.000 (54.000, 69.000)**
- Min - Max	50.000 - 79.000	45.000 - 79.000	**45.000 - 79.000**
BMI			
- Mean (SD)	30.805 (6.060)	28.449 (4.760)	**29.339 (5.410)**
- Median (Q1, Q3)	30.000 (26.690, 33.780)	28.100 (25.000, 31.500)	**28.800 (25.500, 32.395)**
- Min - Max	16.720 - 71.910	16.900 - 48.700	**16.720 - 71.910**
Gender	1476 (60.8%)	2331 (58.2%)	**3807 (59.2%)**
knee_stiff_day_limit			
- Mean (SD)	4.693 (9.692)	3.327 (7.724)	**3.842 (8.545)**
- Median (Q1, Q3)	0.000 (0.000, 2.000)	0.000 (0.000, 1.250)	**0.000 (0.000, 2.000)**
- Min - Max	0.000 - 30.000	0.000 - 30.000	**0.000 - 30.000**
diff_upstr	2058 (84.8%)	2161 (54.0%)	**4219 (65.6%)**

KPACT30	2152 (88.7%)	1031 (25.7%)	**3183 (49.5%)**
Hist_surg	536 (22.1%)	898 (22.4%)	**1434 (22.3%)**
back_pain	1856 (76.5%)	2278 (56.9%)	**4134 (64.3%)**
BP_freq_30			
- 0	571 (23.5%)	1726 (43.1%)	**2297 (35.7%)**
- 1	116 (4.8%)	548 (13.7%)	**664 (10.3%)**
- 2	361 (14.9%)	1104 (27.6%)	**1465 (22.8%)**
- 3	1080 (44.5%)	399 (10.0%)	**1479 (23.0%)**
- 4	295 (12.2%)	227 (5.7%)	**522 (8.1%)**
- 5	4 (0.2%)	0 (0.0%)	**4 (0.1%)**
Lim_act_bp_30	420 (17.3%)	574 (14.3%)	**994 (15.5%)**
HRT_PROB	145 (6.0%)	128 (3.2%)	**273 (4.2%)**
DIABETES			
- 0	2129 (87.7%)	3718 (92.9%)	**5847 (90.9%)**
- 1	257 (10.6%)	286 (7.1%)	**543 (8.4%)**
- 8	41 (1.7%)	0 (0.0%)	**41 (0.6%)**
COPD			
- 0	2287 (94.2%)	3924 (98.0%)	**6211 (96.6%)**
- 1	98 (4.0%)	80 (2.0%)	**178 (2.8%)**
- 8	42 (1.7%)	0 (0.0%)	**42 (0.7%)**
ULCER	136 (5.6%)	94 (2.3%)	**230 (3.6%)**
STROKE	97 (4.0%)	111 (2.8%)	**208 (3.2%)**
ASTHMA	201 (8.3%)	335 (8.4%)	**536 (8.3%)**
DEPRESSION			
- Mean (SD)	7.740 (7.768)	6.165 (6.506)	**6.759 (7.050)**
- Median (Q1, Q3)	6.000 (2.000, 11.000)	4.000 (2.000, 9.000)	**5.000 (2.000, 9.000)**
- Min - Max	0.000 - 58.000	0.000 - 57.000	**0.000 - 58.000**
WOMAC			
- Mean (SD)	22.785 (18.191)	14.859 (15.911)	**17.851 (17.240)**
- Median (Q1, Q3)	20.000 (7.000, 35.000)	9.000 (2.000, 23.000)	**13.000 (3.000, 28.400)**
- Min - Max	0.000 - 96.000	0.000 - 96.000	**0.000 - 96.000**
KL_score	**1288 (53.1%)**	**1773 (44.3%)**	**3061 (47.6%)**

There is a discrepancy in the number of subjects for the original and extended analysis in both cases. This is due to the level of missingness present in the data. For the extended dataset, the data sizes are those for which missing values have been removed. Some

variables were excluded from the analysis due to the high levels of missingness present in the data. These included the original variable of whether baseline symptoms were present, and the potential variables of interest - the presence of kidney disease and the smoker status. Removing these allowed a large enough dataset to remain that could be used in the modelling and subsequent analysis.

For this study, the best method for dealing with the missing values was to use a complete case analysis. There has been two approaches used to accomplish this. First is to remove variables that contain a large amount of missing values. By removing variables, it is possible to preserve the size of the remaining data. The other approach is to remove subjects with at least one missing value. By first removing the variables with a high level of missingness and then removing the individuals with missing information, it was possible to maximise the available data.

7.3. Study Design

7.3.1. Specifics of the MTL method used

Multi-task learning is a subfield of machine learning. There are four primary methods: Assistant Task, Piling, Transfer of Knowledge, and Group Online Adaptive Learning. All four of these approaches take multiple learning tasks at once while exploiting the commonalities and differences across the different tasks [247]. In other domains when this approach has been used, the use of combined datasets on standard machine-learning models has improved learning efficiency in the models and prediction accuracy. As this method makes use of sharing information, it is particularly useful in areas where the tasks are undersampled [248]. The most commonly used methods for this sort of data are the Assistant Task and Piling approaches.

MTL makes use of datasets that are similar to add additional available information to the model with the aim of improving model performance. This type of model trains multiple datasets at once, helping the model learn from multiple sources. This type of learning is useful as many clinical datasets are small, which typically makes them not suitable for machine learning. By doing this, the data is then made suitable for applying machine learning models. This is only the case if the datasets are in some way related and the newly added datasets add information to the problem. Figure 7-1 shows crudely how data from each dataset can be used when training the models.

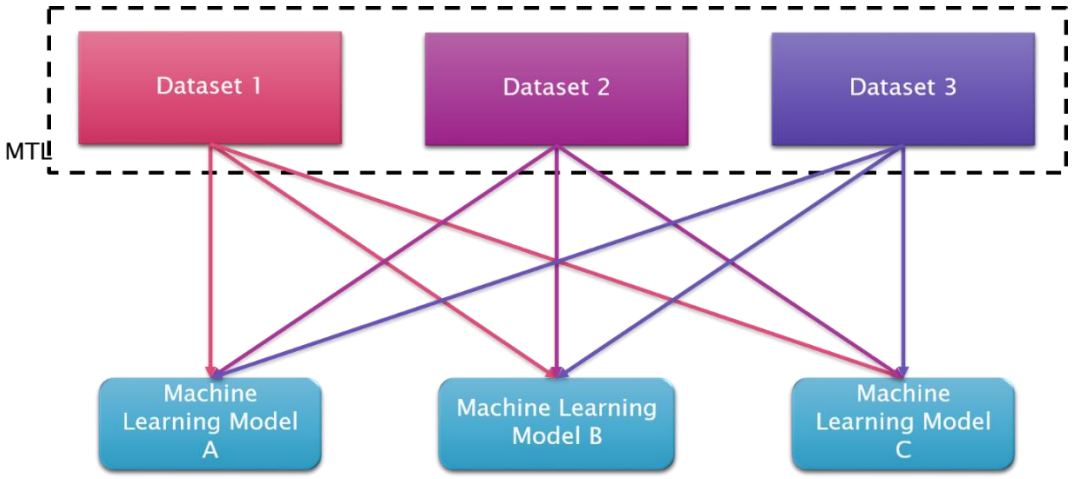


Figure 7-1: This is a graphic showing how a model can take information from multiple sources. This process is known as Multi-Task Learning.

The type of MTL used in this analysis is the Piling approach, shown in Figure 7-2. The piling approach takes all of the datasets to be investigated and 'piles' them up, making one large dataset. The resultant dataset is then used to train the model so that the model is not trained to overfit to the original training data.

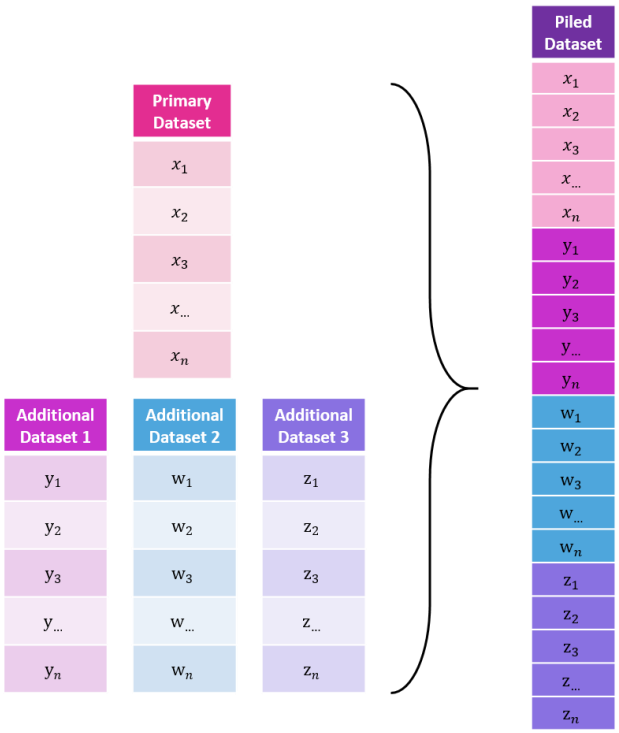


Figure 7-2: The diagram shows how Piling MTL works. The datasets to be used in the analysis are all stacked on top of one another, with only the primary dataset being used for testing purposes.

To better understand the piling approach for this given problem Figure 7-3 builds on the foundations of the idea shown in Figure 7-2.

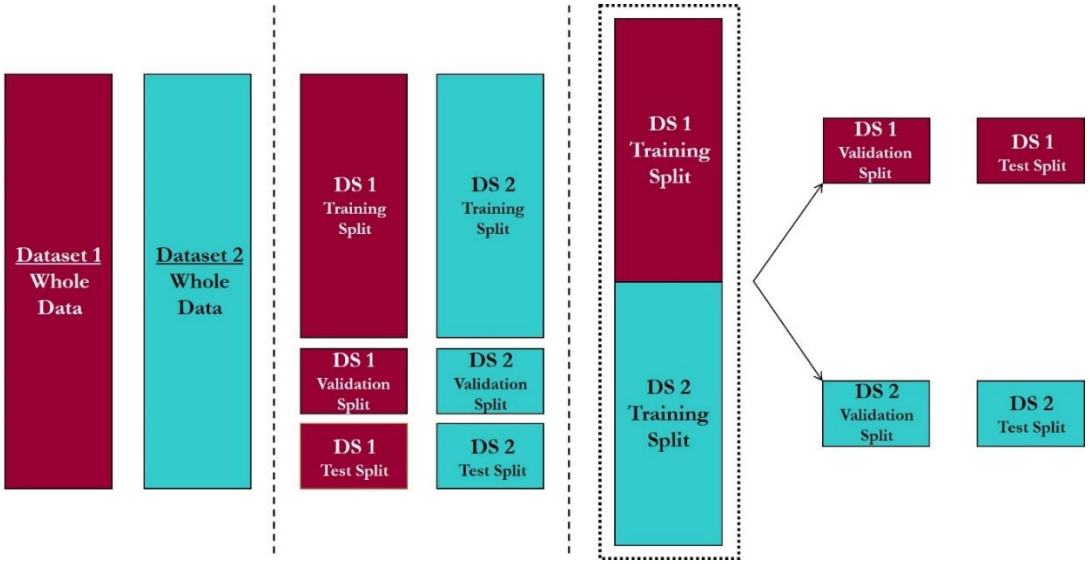


Figure 7-3: Diagram to represent how the data is split for MTL and where each data split is used.

In Figure 7-3 we can see the way the data is taken from a single dataset and transformed into a useable format for MTL. Step one is to take the original datasets, in this analysis OAI and MOST, and split those into training, validation and test splits. As this approach of MTL is piling, the next step is to pile the two training samples. This new training set is used to train the models. This is where the learning from multiple sources takes place. The next step is to consider only one of the datasets for assessing model performance, either dataset 1 or dataset 2. This step involves using the validation and test set from the chosen data to assess the performance for predicting on that chosen dataset only. This ensures that the model is focusing on either dataset 1 or dataset 2 when considering performance. This is key, as the aim is to optimise the models performance.

7.3.2. Specifics of the Neural Networks Applied in Analysis

In Chapter 3, different approaches were utilised when trying to diagnose KOA from the data. That analysis found logistic regression to be the optimal approach, in terms of performance and interpretability. In applying MTL to the problem logistic regression is used as a baseline comparison and neural networks were the main focus. Neural networks can take any number of architectures and to determine if the improvement was because of the neural network or the MTL implementation different architectures have been tested.

The networks used in the analysis in this chapter consist of fully connected networks made up of a single layer with either four, five or ten nodes, or two layers with the first containing twelve nodes and the second containing seven nodes. The architecture for the two-layer network and the single layer with five nodes is shown in Figure 7-4. By taking

this approach, the complexity is varied to determine what, if any improvement can be attributed to the MTL implementation.

The networks were also tested using no control measures, a hidden dropout layer and a visible dropout layer, with the dropout rate set at intervals of 0.1 from 0.2 to 0.5. The dropout layers were implemented to help control for overfitting [249]. The values for the dropout were set as such to provide some effect on the network, as a probability too low would have had minimal effect, and a value greater than 0.5 would result in under-learning of the network.

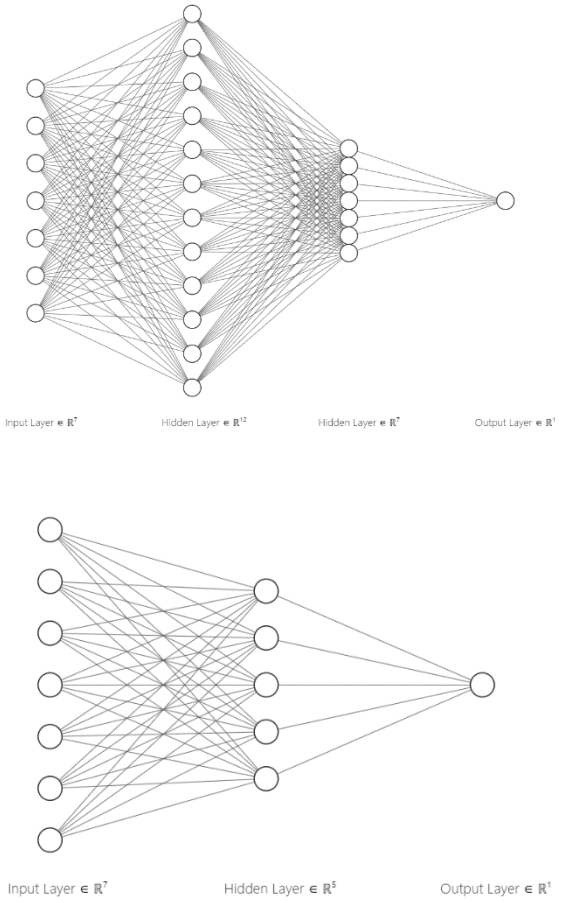


Figure 7-4: A visual representation of two of the four architectures used in this analysis. The network on the top panel is the two-layer network with 12 nodes on the first layer and 7 on the second, and the network on the bottom panek is the single layer network with 5 nodes on the hidden layer. This is varied by having either 4 or 12 nodes in the other models.

The neural networks all used the binary cross-entropy loss function and the Adam optimizer. The binary cross-entropy loss function will produce high values for poor predictions and low values when a prediction is of good quality. A poor prediction would be where the data suggests the probability of belonging to the positive class is low, and this would require a loss to penalize this value. The loss function is defined in Equation

7-1. The optimizer, Adam, used in these networks is an extension to stochastic gradient descent methods used to update network weights iteratively on the training data. This optimiser is both computationally efficient and easy to implement which is a requirement for networks dealing with vast amounts of data.

$$H_p(q) = -\frac{1}{N}\sum_{i=1}^{N} y_i \cdot \log\big(p(y_i)\big) + (1 - y_i) \cdot \log\big(1 - p(y_i)\big)$$

Equation 7-1: The formula for Binary Cross-Entropy.

Where: y ~ the label in the data, in this case it is the knee osteoarthritis status,

$p(y)$ ~ The predicted probability if KOA for all N subjects.

In the neural network implementation used in this analysis there are two different activation functions used. The hidden layers contain the ReLU activation function, whilst the output layer uses the sigmoid activation function.

The ReLU, or rectified linear unit, function is the most used activation function in the world as of 2017 [250]. ReLU enables better training of deeper network models [251] meaning that it often achieves better model performance. It belongs to the family of ridge activation functions that are multivariate functions acting on linear combinations of the input variables, shown in Equation 7-2. The ReLU function is a piecewise linear function that will output the input directly if it is positive, otherwise the output will be zero, shown in Equation 7-3. Because ReLU are nearly linear, they preserve properties that make linear models easy to optimise. They also preserve many of the properties that make linear models generalise well.

$$\phi(\boldsymbol{v}) = \max(0, a + \boldsymbol{v}'\boldsymbol{b})$$

Equation 7-2: The standard presentation of the ReLU activation function.

$$f(x) = \begin{cases} 0 \text{ for } x < 0 \\ x \text{ for } x \geq 0 \end{cases}$$

Equation 7-3: Piecewise representation of the ReLU function.

The sigmoid activation function is used on the output layer. This function is typically used for models where predicting an output as a probability. Sigmoid activation functions are especially useful when the predictions are of a binary class, such as that in this analysis. The outcome will always only exist between zero and one. The formula is shown in Equation 7-4.

$$\sigma(\boldsymbol{z}) = \frac{1}{1 + e^{-z}}$$

Equation 7-4: The equation of the sigmoid activation function.

The neural networks in this analysis, as will all, have three types of layers: input, hidden and output. The input layer contains the same number of neurons as the number of features present in the data. Next are the hidden layers. One hidden layer is sufficient for the vast majority of problems. There are many rules to advise on the configuration of the hidden layers, such as to have a single hidden layer and to have the number of nodes in the hidden layer as the mean of the neurons in the input and output layers. Although these are common guides, there is some speculation about the relevance of the calculation to determine number of neurons in the hidden layer. One thing is certain however, is that if the hidden layer contains too many neurons the model is likely to overfit to the noise in the data. The final layer is the output layer. Every network has exactly one output layer. The network used in this analysis is for binary classification so the output has a single node. If the network were for multi-class classification, the output layer would contain a single node per class label in the model.

7.3.3. Partial Dependency Plots

Explainability is a priority, more so as more responsibility is put on machine learning in everyday situations. Models decide if emails are classified as spam, or if a loan application is approved so understanding and being able to explain the model making these decisions is ever more important. This is also true for medical applications of machine learning approaches.

By being able to explain a model and provide insight, there is more chance of trust and usability in the models created. Models like logistic regression are easily interpreted and explainable. This is one of the reasons they are used in many application domains. At the other end of the scale there are black box models such as neural networks which have difficult to comprehend steps between the inputs and the outputs in the model. A scale of accuracy vs explainability is shown in Figure 7-5.

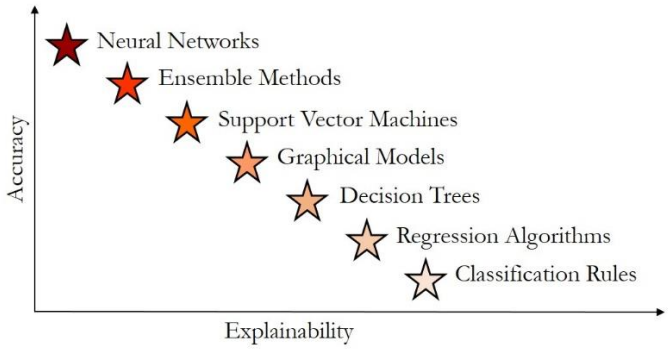


Figure 7-5: A visual representation of the trade-off between model performance and explainability. Traditionally, the more explainable the model, the less accurate the results.

Partial dependence shows how a certain feature can influence a prediction. By fixing all bar one feature, it is possible to see how that given feature influences the outcome [252]. These kind of insights allow the predictions to be thought of intuitively and can enable the users to make sense of the models, which likely leads to more trust in the models predictions.

Let X_S be the set of input features of interest, and X_C be the complement. The partial dependence of the response f at a point x_S is:

$$pd_{X_S}(x_S) \stackrel{\text{def}}{=} \mathbb{E}_{X_C}[f(x_S, X_C)]$$
$$= \int f(x_S, x_C) p(x_C)\, dx_C,$$

Where $f(x_S, x_C)$ is the response function for a given sample whose values are defined by x_S for the features in X_S and x_C for the features in X_C.

The model will change the value of a given variable each time a prediction is made. By keeping a record of the predictions, it is possible to see how this variable affects the overall prediction. This is repeated and average predictions for each point are calculated. These averages are then used for the partial dependency plots. The partial dependence method also allows consideration for feature interactions.

The plot allows visualisation of this interaction, further increasing the interpretability in the model. The visual representations provided by use of the partial dependency plots helps to deliver insight and simple interpretations to more complex concepts. This distilling of information from complex abstract ideas to logical insights helps to make more complex machine-learning methods, such as neural networks, easier for a non-

technical audience to understand. This is one way to help increase the chance for models to be utilised in areas where human understanding and explainability are paramount.

The partial dependence applies prescriptive analytics to conventional machine learning approaches. This helps to ensure that the information in the data is deeply understood, and allows this information to be turned into actionable insights. Doing this adds value to the analysis by making it accessible to those without expert insight into the methods used.

7.3.4. Justification for Analysis Approach

Using different data from both the OAI and MOST studies allows the incorporation of the MTL modelling to not only answer the question of what features best predict the incidence of KOA but is MTL useful in this instance for improving model performance. By using an extended pool of data features there is the potential for the MTL approach to uncover different relationships in a subjects profile and their likelihood of having KOA.

If the MTL model resulted in higher model performance this would have the potential to be used by clinicians as the model could be optimised by increasing the pool of data used train the model on. This could also open the door to examining other features that could be early indicators of disease presence.

The other justification for using this modelling approach to the analysis is to also incorporate the key ideas of interpretability and explainability into an approach that is typically viewed as a black box. By incorporating the partial dependency plots and the neural networks alongside the MTL implementation there is real potential to develop high performing diagnostic models that are easily explainable. This would provide the necessary first step in developing guidelines for implementation into a clinical environment.

Having several different network architectures used in the analysis ensures that different combinations are considered when looking for the optimal model, and that the true optimal for a set of conditions is found and the resulting findings are not chance results. The architectures considered look at two different data combinations, using four different networks with three configurations for controlling overfitting, with two subject to a further four possible drop-out rates, shown in Figure 7-6.

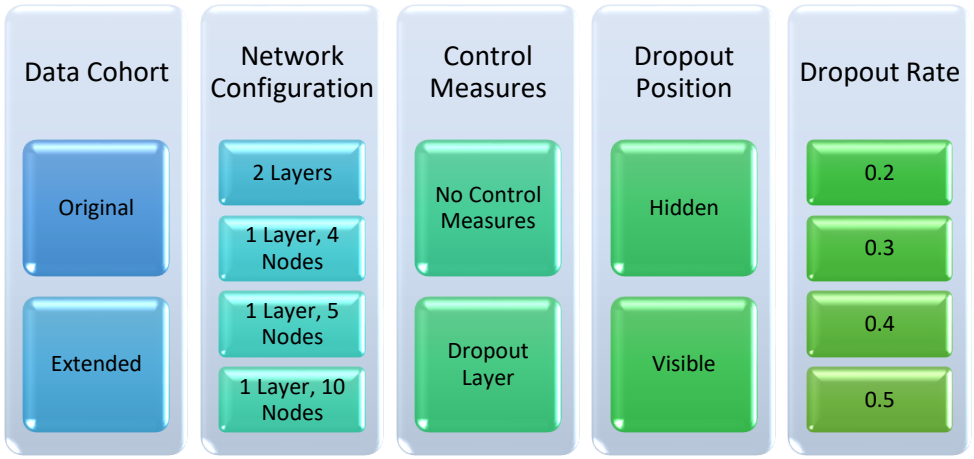


Figure 7-6: Depiction of choices for the model architecture process.

7.4. Results from Analysis

Running the models for the different configurations of neural network produced a total of 864 model performances. To determine the best performing models the test AUC for each dataset and strategy has been put into descending order and the top one for each selected. This is important, as the best performing model for each configuration will have partial dependencies applied to them in order to visualise how each feature contributes to the overall chance of having KOA at the baseline visit with a clinician. In this section, only the partial dependence plots for selected interactions in a variety of the datasets will be considered.

Table 7-3: A table showing the highest performing single and multi-task learning model for each configuration of the neural network.

Data Set used	Learning Type	Network Configuration	Test AUC (95% CI)
OAI	Single Task Learning	1 layer, 10 nodes, hidden dropout 0.3, batch 50	0.7662 (0.7384 – 0.7941)
	Multi-task Learning	1 layer, 5 nodes, hidden dropout 0.2, batch 100	0.7611 (0.7329 – 0.7893)
MOST	Single Task Learning	1 layer, 5 nodes, batch 75	0.7598 (0.7145 – 0.8051)
	Multi-task Learning	1 layer, 5 nodes, batch 75	0.7533 (0.7070 – 0.7996)
OAI Extended	Single Task Learning	1 layer, 5 nodes, visible dropout 0.2, batch 50	0.7316 (0.7003 – 0.7630)
	Multi-task Learning	2 layers, 12/7 nodes, hidden dropout 0.3, batch 50	0.7399 (0.7090 – 0.7708)
MOST Extended	Single Task Learning	1 layer, 5 nodes, batch 50	0.8069 (0.7728 – 0.8410)
	Multi-task Learning	1 layer, 10 nodes, hidden dropout 0.2, batch 100	0.8167 (0.7837 – 0.8497)

From looking at the information in Table 7-3 the test AUC for both OAI and MOST when considering the original cohort only, single task learning (STL) performs marginally better then when MTL is used on the same data. This is not a surprising result, as for the OAI model performance with logistic regression from Chapter 3 the test AUC was 0.763, which is comparable to the neural network performance here. The model appears to have reached the 'ceiling' of the data in terms of performance. A similar conclusion can be reached when considering the MOST data for single and multi-task learning. The phenomenon where a model suffers after using multi-task learning, negative transfer, cannot be deemed the cause with certainty as the difference in performance is roughly 0.5% in both cases.

However when considering the extended data, both for OAI and MOST the models perform marginally better on the MTL approach. As this slight performance improvement has been noted in both the OAI and MOST extended data, the MTL application has positively influenced the model. Despite this slight improvement of roughly 1%, this is not significant enough to state this has been caused by the use of MTL.

In each of the bivariate graphs, there are three bars in the plots. The bars on the left side and top of each plot show the range of variable options for each variable. The bar on the right side of the plots give the range of values relating to the outcome, in this case, the presence of KOA in the subject.

The features such as age and BMI change the risk of having KOA at first presentation, with advanced age and high BMI both risk factors for disease. The partial dependency plots showing the interaction between age and BMI for the OAI data with both single and multi-task learning are displayed in Figure 7-7. In both instances, it is clear to see that high BMI above 30 has a more important role in the likelihood that someone will have KOA as the plots have more of a vertical element. Below 30 the BMI and age are nearly equally influential for indicators for disease presence.

Age and BMI are known risk factors for KOA. The variables in the analysis also consider some anecdotal risk factors, such as the number of days in the past month knee stiffness has limited activity. In Figure 7-8 BMI and limiting knee stiffness (knee_stiff_day_limit) are plotted to see the interaction and where the variables have influence in determining if someone has KOA.

In the single task plot knee stiffness is relatively important up to the point a persons BMI goes above 30, and then BMI has increasing influence over the KOA status until a BMI of above 34 where BMI is the only feature influencing the KOA status. In the MTL plot, knee stiffness is less influential when the BMI reaches about 25. A similar pattern from the single task learning for BMI being the sole driver in indicator for KOA is seen with a BMI beyond about 30. The training pool for the MTL model was greater as it incorporated both the OAI and MOST data which could have caused this shift to more sever outcomes at lower BMI ranges.

The results for the models tested on the original MOST data are consistent with those from the OAI test results. However, Figure 7-9 shows the PDP for BMI and limiting knee stiffness, similar to that in Figure 7-8, with obvious differences. As can be expected, in STL the main driver behind KOA status is the BMI. Conversely, in the MTL analysis the main influence switches to limiting knee stiffness. This is likely due to the majority of subjects in the MOST dataset suffering with KOA, therefore influencing the link between limiting knee stiffness and presence of KOA.

The use of an extended dataset offered the chance to consider features that are of clinical relevance but not commonly associated with KOA, such as depression and heart problems. Plots produced using partial dependencies, however, show that no conclusions about the clinical significance of the variables can be drawn based on this analysis. This is due to the PDP showing that presence of KOA remains nearly constant as the individual features vary. This is shown in Figure 7-10.

Figure 7-11 shows the interaction between the WOMAC score and the number of days the subject has experienced limiting knee stiffness. In both cases the WOMAC score is more influential in the overall determination of the presence of KOA. (A) however shows that days experiencing knee stiffness has equal influence in a linear relationship with WOMAC up to a WOMAC score of about 35, at which point the sole influencing feature becomes the WOMAC score. In the MTL plot, (B), a much lower WOMAC score of about 12 becomes the point where sole influence is passed to WOMAC, although in the MTL plot the influence was always lead by WOMAC.

Considering the extended data cohorts from MOST with both STL and MTL Figure 7-12 shows the PDP interactions for BMI and WOMAC. Both STL and MTL appear to have equal contribution from both features up to a BMI of between 32 and 35, when BMI is the main influence. For the STL plot, (A), lower WOMAC scores are more in control for

BMI up to about 30 when the shift begins to take place. In the MTL plot, (B), the relationship appears mainly linear until the higher BMI values are reached.

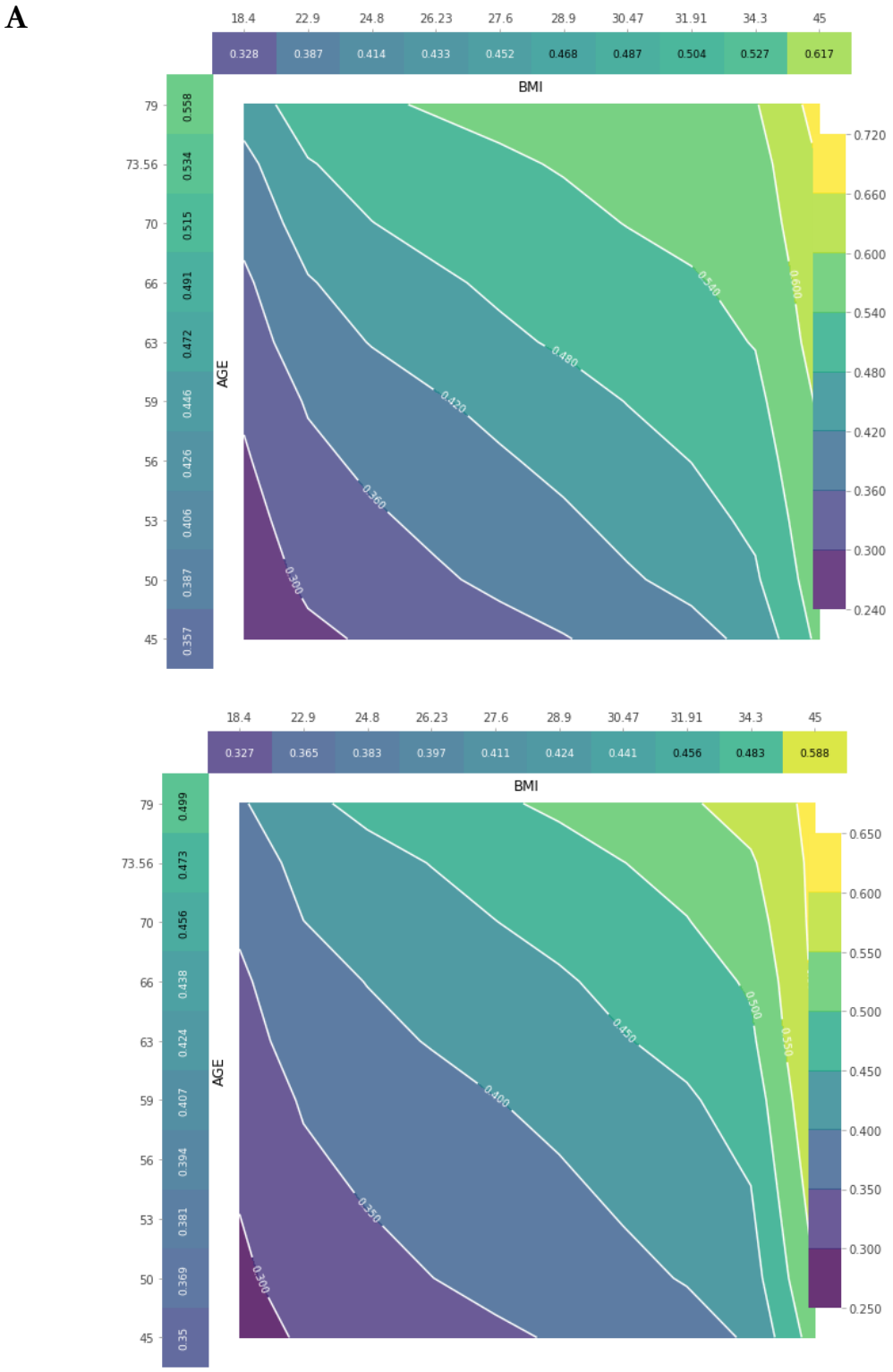
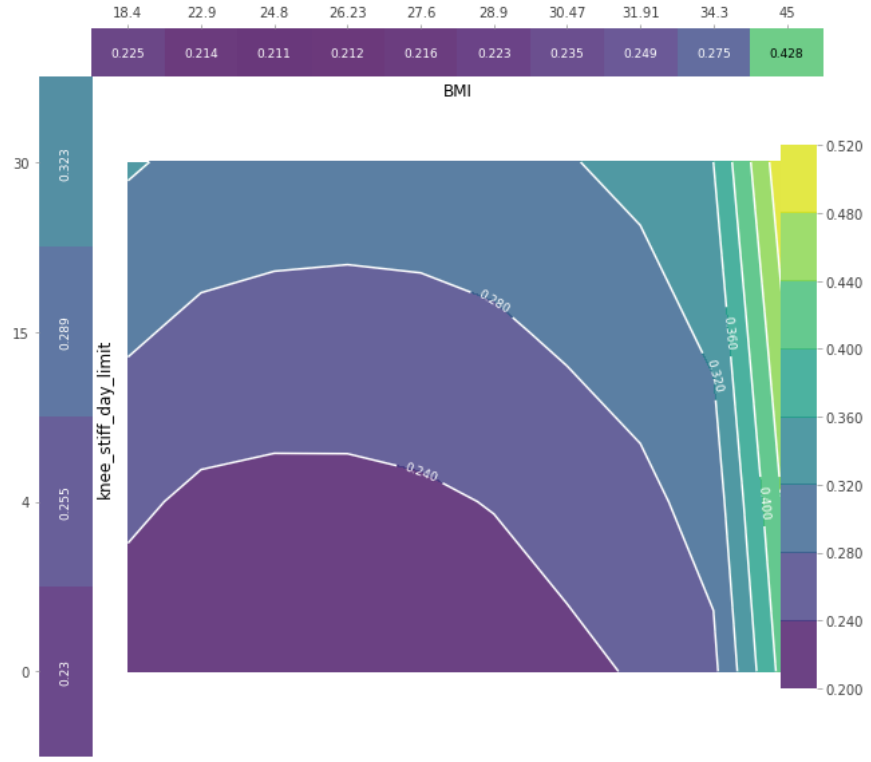


Figure 7-7: Partial dependency plots for Age and BMI. (A) shows the PDP for the single task learning approach on the OAI test data. (B) shows the plot for the MTL approach on the OAI test data after training on both MOST and OAI data.

A



B

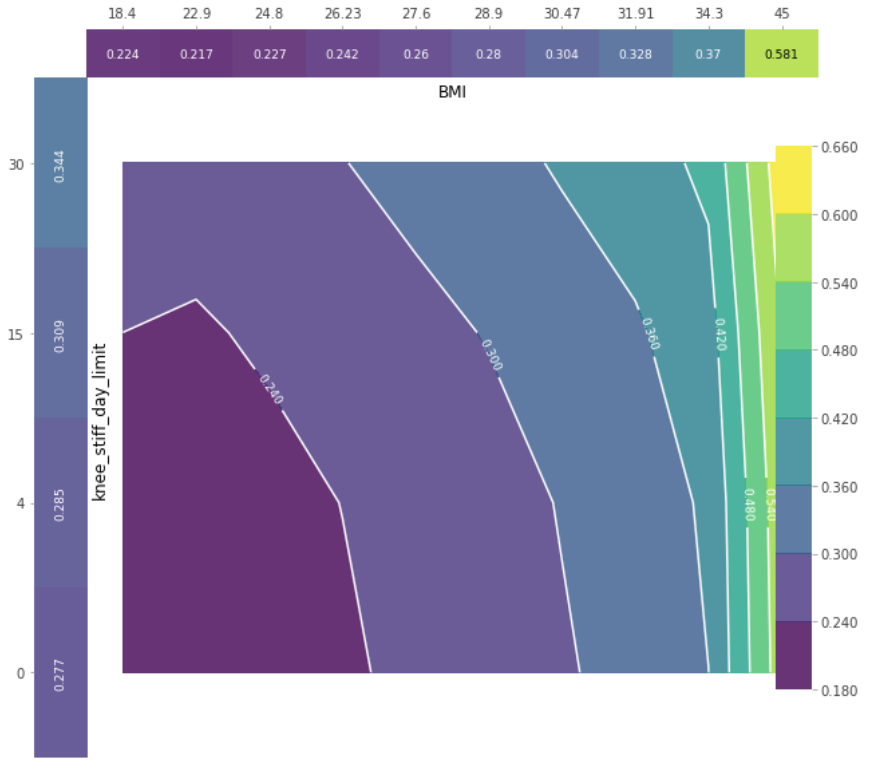


Figure 7-8: Partial dependency plots for Days of knee stiffness limiting activity and BMI. (A) shows the PDP for the single task learning approach on the OAI test data. (B) shows the plot for the MTL approach on the OAI test data after training on both MOST and OAI data.

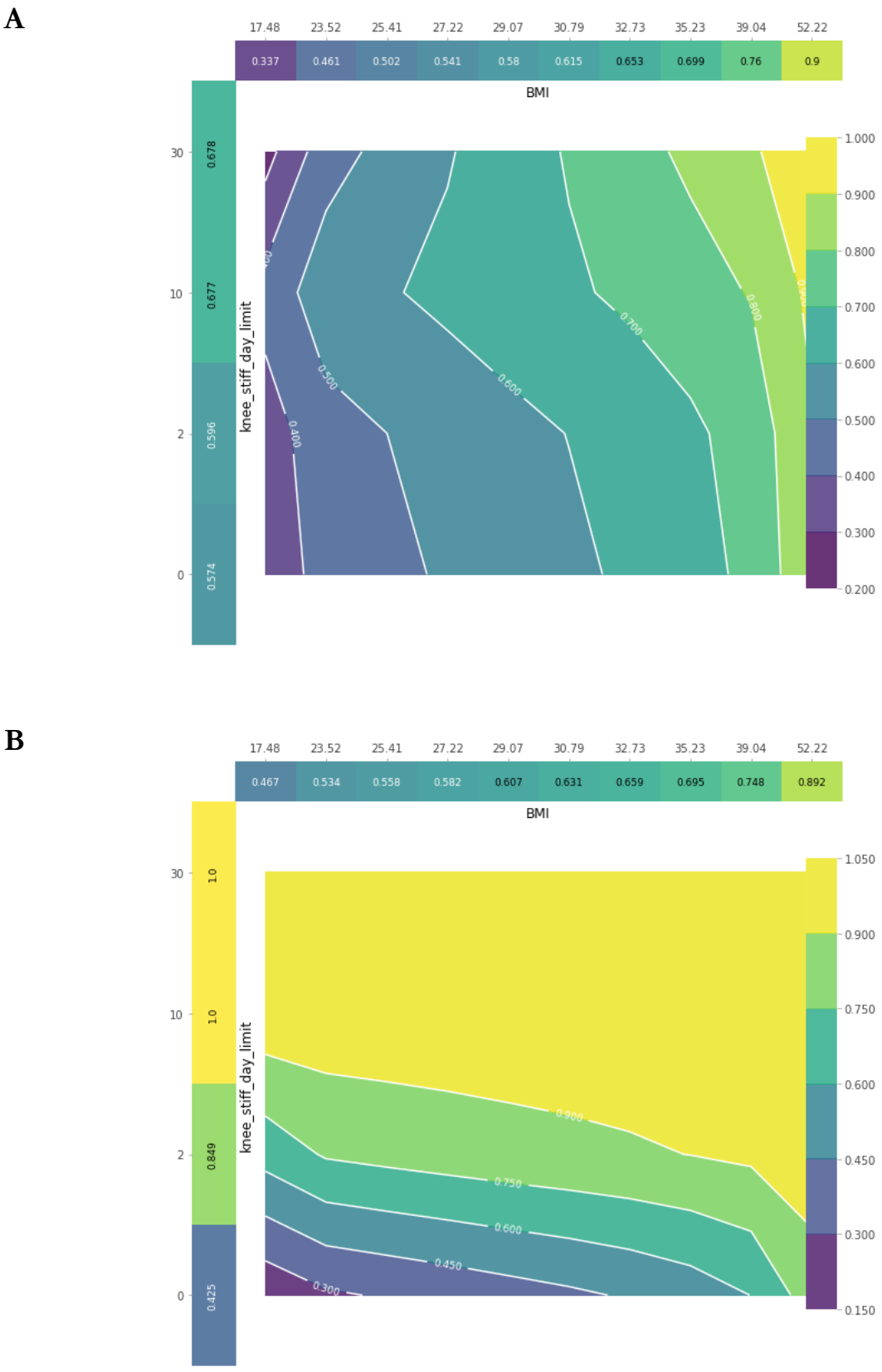
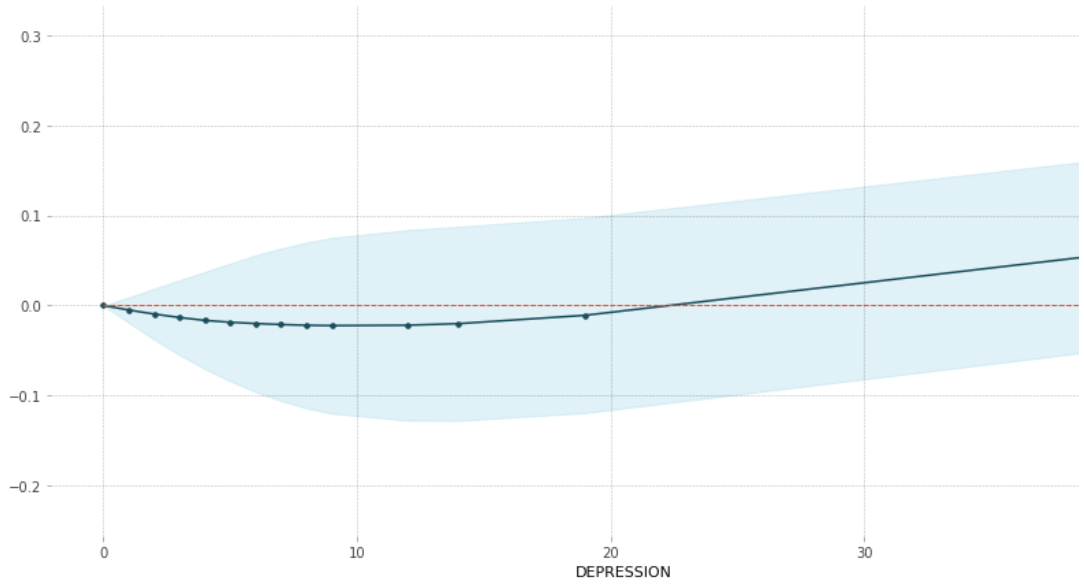


Figure 7-9: Partial dependency plots for Days of knee stiffness limiting activity and BMI. (A) shows the PDP for the single task learning approach on the MOST test data. (B) shows the plot for the MTL approach on the MOST test data after training on both MOST and OAI data.

A



B

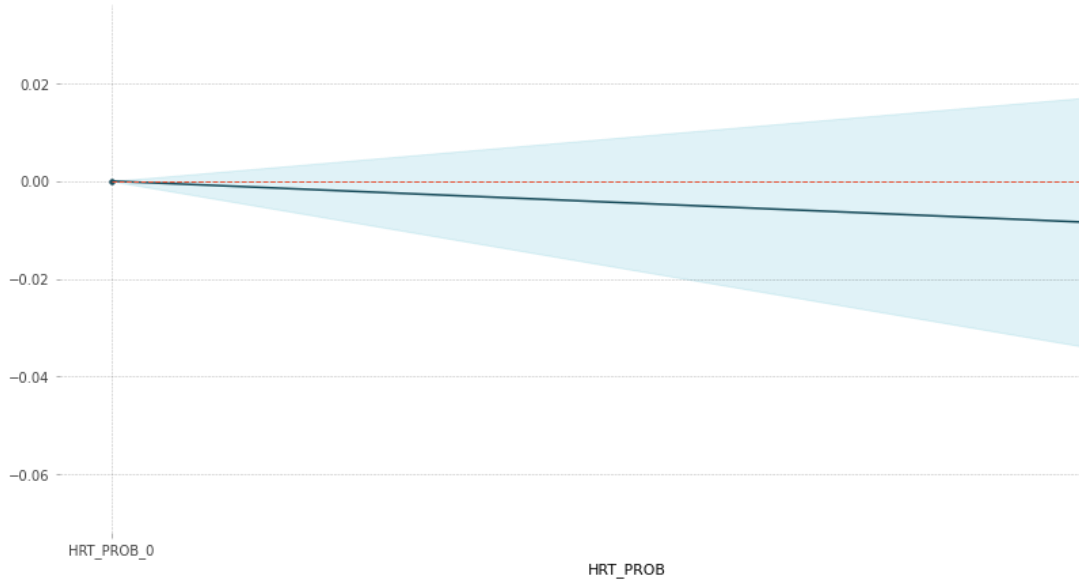
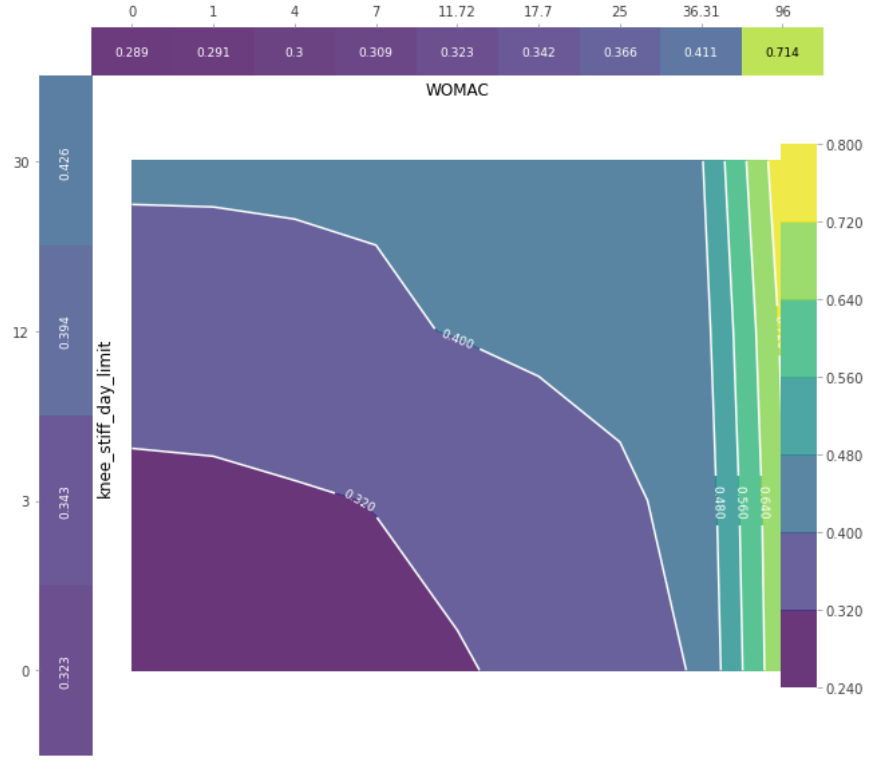


Figure 7-10: Partial dependency plots for Depression and Heart Problems. (A) shows the PDP for depression in relation to the presence of KOA. (B) shows the PDP for the relation to KOA and heart problems. Both of these are generated using the extended OAI data with STL.

A



B

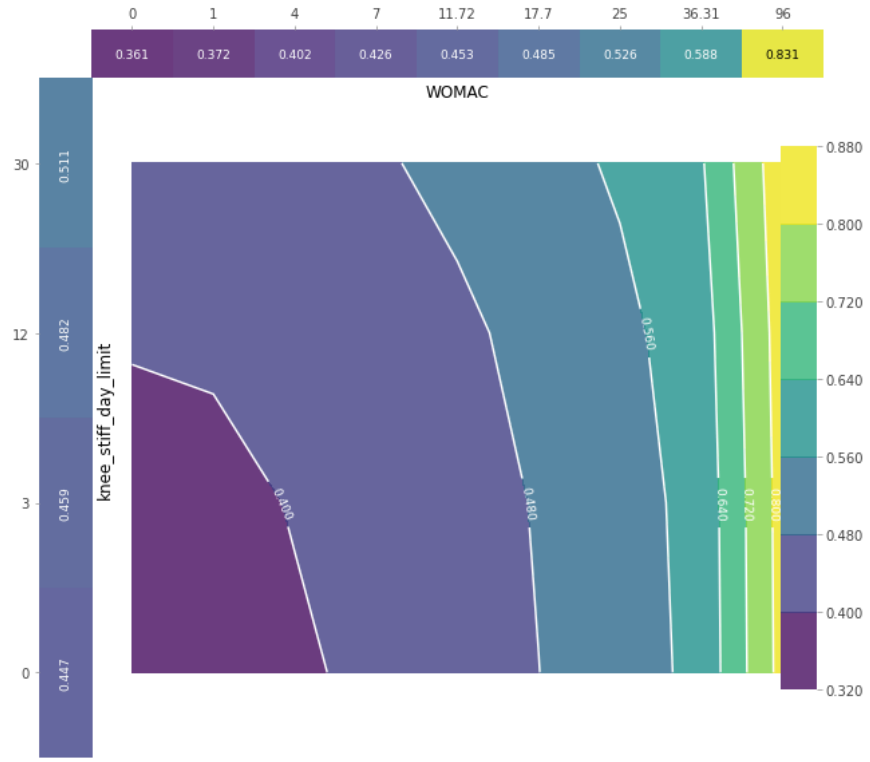
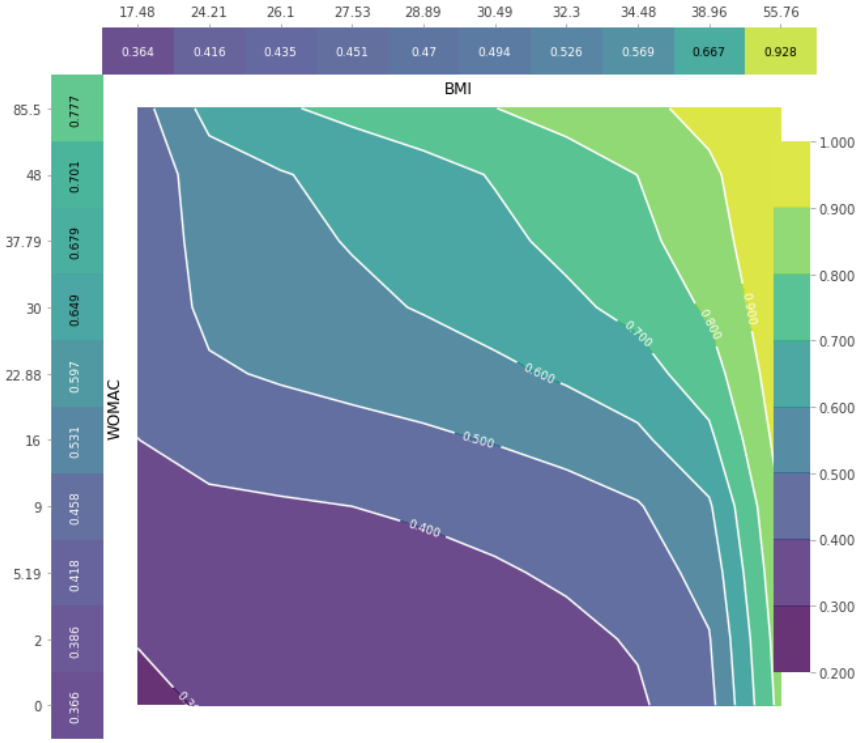


Figure 7-11: Partial dependency plots for WOMAC and Days of knee stiffness limiting activity. (A) shows the PDP for the single task learning approach on the extended OAI test data. (B) shows the plot for the MTL approach on the extended OAI test data after training on both MOST and OAI data.

A



B

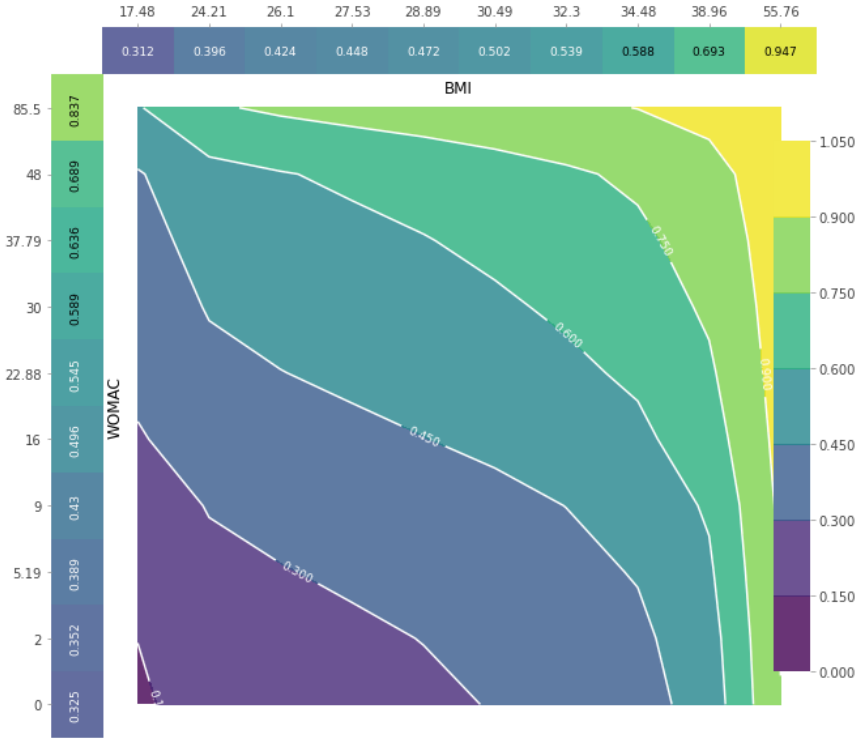


Figure 7-12: Partial dependency plots for WOMAC and BMI. (A) shows the PDP for the single task learning approach on the extended MOST test data. (B) shows the plot for the MTL approach on the extended MOST test data after training on both MOST and OAI data.

7.5. Discussion

By having the two analyses, one with the original pool of variables and one with the extended variable cohort, there can be comparison between which set of data was more suited to the MTL implementation. As a lot of the variables for the extended cohort were binary 'yes/no' answers the information that could be garnered was limited. Therefore, for this data the standard variables provided more insight into the decision boundaries, with the exception of the WOMAC scores, which as a continuous variable provided a glimpse of how the boundary varies in a more gradual way than the severe binary options.

The partial dependence plots provide an easily interpretable visualisation of the way the neural network is making decisions and therefore classifications. This level of insight is required for models that have the potential to be implemented in clinical settings, where explainability is paramount.

As this study was a first step into applying MTL to the problem of diagnosing KOA, there is some evidence to suggest that further tuning of the models, such as hyper-parameter tuning, and implementation of different MTL methods may prove more beneficial to the overall model performance.

The results show, and further make certain, the role that BMI has in the presence of KOA, with majority of the results containing BMI being heavily influenced in that direction. This cements that a higher BMI is an indication that KOA should be in the clinicians mind when discussing any potential KOA related issues. Similarly, age is a known risk factor for the onset of KOA, and this was shown to hold true with the use of the PDP for the OAI test data, specifically illustrated using the original cohort.

The model performance for all four data cohorts used in this chapter in MTL and STL are similar to one another, and contain no significant differences. However, the difference in model performance using the original and extended cohorts for the OAI test set demonstrate a drop in performance, whereas for the MOST test data there is a performance increase. This illustrates that the model, with consistent data, will perform differently depending on the cohort of data used to test. In terms of producing a model with clinical benefit, the focus would be to remove this variability, where the model would perform with relative consistency for the given test case.

Chapter 8: Discussion

8.1. Conclusions

Knee osteoarthritis is a huge cost on the NHS and worldwide [13], [16], [133]. This cost is only going to get worse due to the global ageing population. As the majority of those who suffer from KOA are older, the ageing population will mean that there are more people suffering from KOA [253], [254]. As many of those affected are of advanced age, they may also suffer from other conditions. This compound nature of disease and a battery of symptoms is an indicator for a lower quality of life [111], [121].

Current diagnosis of knee osteoarthritis is subjective, and based on subjects presenting with a certain set of features that fit the criteria, as there is no test for knee osteoarthritis. Known risk factors may influence clinicians to a KOA diagnosis for people presenting with typical clinical presentation, but can result in lengthy waits for those who are atypical in their disease presentation.

Those at risk are often unaware they are at risk until they become symptomatic. Within this thesis, we develop tools for clinical implementation for screening, signposting and education for subjects who fit the demographic of sufferers and those at risk.

Chapter 3 describes and develops the modelling for a system that could be adapted for a clinical screening tool. The features in this model are based on the known features of clinical importance when determining if KOA is present, such as age and BMI, along with novel features, such as limiting knee pain and difficulty getting upstairs, as these help gather more insight from the subject. This produced several models with different implementations to allow for a choice that gives the most interpretable model to carry forward. The result of this was the logistic regression model as it is a commonly used implementation in current clinical practice for different modelling of treatments and diseases [255]. This work formed the basis of the framework used for model validation in Chapter 5 and the look at how gender influences KOA in Chapter 6.

Although Joseph at al. [56] produced a model that is an app, the inputs require information from MRI images and focused on those with none or mild KOA determined from an x-ray. Our study used features that can be gathered solely from the subject in any person aged 45 or over to determine their individual risk of having KOA. This could then be used to determine if the person required further interventions, such as x-rays or MRI scans to definitely confirm this determination.

Current advice to those who are likely to develop KOA is to lose weight and move more, which many people find unhelpful as this fails to acknowledge other features that a person may have in their life that may influence the chance of having KOA. The work in Chapter 4 uses time-to-event analysis to calculate if a person is at high or low risk for developing KOA in the next 5 years. This can be used as a tool for patient education to show the influence each feature has on the overall likelihood of developing KOA in the next 5 years based on information gathered at a clinical visit, helping to not only educate, but also improve the advice that is being given out.

For the discrete time analysis of the data presented in Chapter 4, the separation of high and low risk cohorts due to stratification, for both the training and test data is clear. The smoothness of the discrete time fits and the differences with the predictions for continuous time are likely caused by the interval censoring which is better taken into account by the use of discrete time intervals. However, for the work presented in Chapter 4 due to the cluster-time-to-event outcomes, continuous survival analysis is as appropriate as discrete time analysis and the results are equivalent.

The data gathered as part of the Horizon 2020 project, OActive, had limited use in the scope of the diagnostic and predictive modelling due to issues relating to variable agreement with different datasets. However, in Chapter 5 some limited model validation was carried out on the diagnostic model using the OActive data. The work in Chapter 5 also provided external validation from MOST on the OAI developed models for diagnosis and time to development of KOA in a 5-year window.

Chapter 5 used the work from both Chapters 3 and 4 and validates these and builds web apps that could be easily deployed for use in clinical settings as both a clinical decision support and patient education resource. The diagnostic model gives a probability of a person with a given set of features having KOA, whilst the predictive model indicates, based on a set of features, if a person is at high or low risk for developing KOA in the next 5 years.

The extended models produced in Chapters 6 and 7 build on the work from Chapter 3 and look with different clinical questions. Chapter 6 looks to determine if there is any benefit from modelling KOA differently in male and female subjects, including gender specific features. Although there is not a statistically significant difference between the original model proposed in Chapter 3, the inclusion of gender specific features offers a different way to explain what is happening in the model as some generic variables may

have masked the effect of specific features within the output to achieve a nearly equal model performance. By allowing a more granular model to be used, and including additional features relating to gender, there is the potential for a deeper understanding behind the relationship between gender and KOA to be developed. This development could also have far reaching impacts, such as a more personalised evaluation of the patient from step one in the diagnosis process [256].

In Chapter 7, we aimed to utilise MTL as a way of boosting model performance. The results, as expected, did not garner much if any improvement. This may have been as a result of the datasets used, OAI and MOST, being different to allow for information from one to enhance the models capability. In this work, there was also a preliminary look at the effect of including more features gathered from conditions that are unrelated to KOA. The extended modelling produced models with better performance however, these would require further analysis to determine if they were of clinical significance.

In summary, this thesis aims to develop models that can be used to diagnose and predict the likelihood a person has KOA based on demographic information. We provide two externally validated models that are optimised for explainability. This is the key novelty in this study. The resulting models can easily be used and interpreted by a clinician.

8.2. Future Work

While the results of this research are promising, with additional time and resources, the utility of the model predictions could clearly be improved. Work in Chapter 3 could be extended by considering different combinations of variables to be used in the analysis. There is also the potential to incorporate multiclass classification into the problem to determine the probability of having a certain KL grade based on clinical features.

For the OActive data to be included, further experiments and optimisation of the MTL framework are required. An example could be to develop a neural network that has task specific variables built in to be able to provide the data available to the user and produce a result with better suitability to the task in question.

In order to adapt these models to the UK demographic a dataset from the UK would need to be used. This would then allow for validation of the models on data from the UK, helping to cement the usability for these models in Britain. This is because the demographics of the population vary between the US, where this data was gathered, and

the UK where the model would be used. This optimisation could be conducted with the use of MTL to also incorporate the data from the OAI and MOST datasets.

Although the work in this thesis solely relied on clinical data, a next step could be to utilise the information held in images such as x-ray and MRI information to use a multifaceted approach to a prediction model. By using more than one type of data, the multi-task learning approach discussed in Chapter 7 could be further expanded to also include multisource data. This addition to the modelling could help improve model performance by incorporating features contained within images and the clinical data to provide new insight into the connection between medical imaging and symptoms that may have otherwise gone unnoticed.

By creating a screening system for people at risk of developing KOA there is the potential to not only save money but to enhance the quality of life for those who have developed and are likely to develop KOA in the future.

References

[1] C. Giannaki, "Horizon 2020 Call : H2020-SC1-2016-2017 (Personalised Medicine)," 2017.

[2] G. Peat, R. Mccarney, and P. Croft, "Knee pain and osteoarthritis in older adults: a review of community burden and current use of primary health care," *Ann. Rheum. Dis.*, vol. 60, no. 2, pp. 91–97, Feb. 2001.

[3] M. D. Kohn, A. A. Sassoon, and N. D. Fernando, "IN BRIEF Classifications in Brief Kellgren-Lawrence Classification of Osteoarthritis," *Clin. Orthop. Relat. Res.*, vol. 474, 1999.

[4] L. . Jones, D. Golan, S. . Hanna, and M. Ramachandran, "Artificial intelligence, machine learning and the evolution of healthcare," *Bone Joint Res.*, vol. 7, no. 3, pp. 223–225, 2017.

[5] J. Hee Ryu *et al.*, "Measurement of MMP Activity in Synovial Fluid in Cases of Osteoarthritis and Acute Inflammatory Conditions of the Knee Joints Using a Fluorogenic Peptide Probe-Immobilized Diagnostic Kit," *Theranostics*, vol. 2012, no. 2, pp. 198–206, 2012.

[6] R. W. Moskowitz, "The burden of osteoarthritis: clinical and quality-of-life issues.," *Am. J. Manag. Care*, vol. 15, no. 8 Suppl, pp. S223-9, 2009.

[7] J.-P. Lepine and M. Briley, "The epidemiology of pain in depression," *Hum. Psychopharmacol. Clin. Exp.*, vol. 19, no. S1, pp. S3–S7, Oct. 2004.

[8] VersusArthritis, "The State of Musculoskeletal Health 2019," 2019.

[9] GOV.uk, "Methods, data and definitions - GOV.UK," 2018. [Online]. Available: https://www.gov.uk/government/publications/health-profile-for-england-2018/methods-data-and-definitions#quality-of-life-score-eq-5d. [Accessed: 09-Mar-2022].

[10] PHE, "Chapter 3: trends in morbidity and risk factors - GOV.UK," *Public Health England*, 2018. [Online]. Available: https://www.gov.uk/government/publications/health-profile-for-england-2018/chapter-3-trends-in-morbidity-and-risk-factors. [Accessed: 17-Feb-2020].

[11] R. D. Altman, "Early management of osteoarthritis.," *Am. J. Manag. Care*, vol. 16 Suppl M, pp. S41-7, Mar. 2010.

[12] F. Xie *et al.*, "A Study on Indirect and Intangible Costs for Patients with Knee Osteoarthritis in Singapore," vol. 2, no. Supplement 1, pp. S84–S90, 2008.

[13] A. D. Woolf and B. Pfleger, "Burden of major musculoskeletal conditions," *Bull. World Health Organ.*, vol. 81, pp. 646–656, 2003.

[14] A. Chen, C. Gupte, K. Akhtar, P. Smith, and J. Cobb, "The Global Economic Cost of Osteoarthritis: How the UK Compares," *Arthritis*, vol. 2012, pp. 1–6, Oct. 2012.

[15] D. Ellams *et al.*, "8 th Annual Report National Joint Registry for England and Wales Healthcare Quality Improvement Partnership NJR RCC Network Representatives National Joint Registry for England and Wales 8 th Annual Report," 2011.

[16] P. Hamilton, M. Lemon, and R. Field, "COST OF TOTAL HIP AND KNEE ARTHROPLASTY IN THE UK. A COMPARISON WITH THE CURRENT REIMBURSEMENT SYSTEM IN THE NHS.," *Orthop. Proc.*, vol. 91-B, no. SUPP_I, p. 112, 2009.

[17] NICE, "National Clinical Guideline Centre," 2014.

[18] J. Puig-Junoy and A. Ruiz Zamora, "Socio-economic costs of osteoarthritis: A systematic review of cost-of-illness studies," *Seminars in Arthritis and Rheumatism*.

2015.
[19] R. Bitton, "The Economic Burden of Osteoarthritis," *Am. J. Manag. Care*, vol. 15, pp. S230-5, 2009.
[20] R. Myers, "'Invisible' arthritis to cost UK economy £3.43 billion a year through sufferers taking time off work - Mirror Online," *Mirror*, 2017. [Online]. Available: https://www.mirror.co.uk/news/uk-news/invisible-arthritis-cost-uk-economy-11027237. [Accessed: 17-Feb-2020].
[21] C. M. Torio and B. J. Moore, "National Inpatient Hospital Costs: The Most Expensive Conditions by Payer, 2013 #204," 2016.
[22] C. Thomas, B. Ellis, N. Ali, and J. Connor, "Physical activity and musculoskeletal health," 2016.
[23] J. O'Malley, "House of Lords Science and Technology Committee inquiry.," 2017.
[24] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016.
[25] Y. Bengio, "Learning Deep Architectures for AI," *Found. Trends® Mach. Learn.*, vol. 2, no. 1, pp. 1–127, Nov. 2009.
[26] W. Samek, G. Montavon, A. Vedaldi, L. K. Hansen, and K.-R. Müller, Eds., *Explainable AI: interpreting, explaining and visualizing deep learning*. Springer, 2019.
[27] B. Goodman and S. Flaxman, "European Union regulations on algorithmic decision-making and a 'right to explanation,'" 2016.
[28] D. Bacciu, B. Biggio, P. Lisboa, J. D. Martin, L. Oneto, and A. Vellido, "Societal Issues in Machine Learning: When Learning from Data is Not Enough," in *27th European Symposium on Artificial Neural Networks, ESANN, 2019, Bruges, Belgium, April 24-26, 2019*, 2019.
[29] I. Corp., "IBM SPSS Statistics for Windows." IBM Corp., Armonk. NY, 2016.
[30] T. A. Etchells and P. J. G. Lisboa, "Orthogonal search-based rule extraction (OSRE) for trained neural networks: a practical and efficient approach," *IEEE Trans. Neural Networks*, vol. 17, no. 2, pp. 374–384, 2006.
[31] P. J. G. Lisboa, "A review of evidence of health benefit from artificial neural networks in medical intervention," *Neural Networks*, vol. 15, pp. 9–37, 2002.
[32] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," *Nat. Mach. Intell.*, vol. 1, pp. 206–215, 2019.
[33] D. Alvarez-Melis and T. S. Jaakkola, "Towards Robust Interpretability with Self-Explaining Neural Networks," in *32nd Conference on Neural Information Processing Systems*, 2018.
[34] B. J. Heard, J. M. Rosvold, M. J. Fritzler, H. El-Gabalawy, J. P. Wiley, and R. J. Krawetz, "97 BLOOD SERUM TO DIAGNOSE OSTEOARTHRITIS-BIOMARKERS AND MACHINE LEARNING," 2014.
[35] A. Tiulpin, J. Thevenot, E. Rahtu, P. Lehenkari, and S. Saarakkala, "Automatic knee osteoarthritis diagnosis from plain radiographs: A deep learning-based approach," *Sci. Rep.*, vol. 8, no. 1, pp. 1–10, 2018.
[36] OActive and D. Tsaopoulos, "Data mining tools for knowledge extraction - OActive Deliverable Report," 2021.
[37] N. Lazzarini *et al.*, "A machine learning approach for the identification of new biomarkers for knee osteoarthritis development in overweight and obese women," *Osteoarthr. Cartil.*, vol. 25, no. 12, pp. 2014–2021, Dec. 2017.
[38] L. Minciullo, P. A. Bromiley, D. T. Felson, and T. F. Cootes, "Indecisive Trees for Classification and Prediction of Knee Osteoarthritis," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 10541 LNCS, pp. 283–290, 2017.

[39] A. Tiulpin *et al.*, "Multimodal Machine Learning-based Knee Osteoarthritis Progression Prediction from Plain Radiographs and Clinical Data," *Sci. Reports 2019 91*, vol. 9, no. 1, pp. 1–11, Dec. 2019.
[40] Q. Kadhim Al-Shayea, "Artificial Neural Networks in Medical Diagnosis," *IJCSI Int. J. Comput. Sci. Issues*, vol. 8, no. 2, pp. 150–154, 2011.
[41] D. Kim *et al.*, "672 RISK PREDICTION OF KNEE OSTEOARTHRITIS USING ARTIFICIAL NEURAL NETWORK," 2014.
[42] G. J. Katuwal and R. Chen, "Machine Learning Model Interpretability for Precision Medicine," 2016.
[43] D. T. Felson, "The epidemiology of knee osteoarthritis: results from the Framingham Osteoarthritis Study.," *Semin. Arthritis Rheum.*, vol. 20, no. 3 Suppl 1, pp. 42–50, Dec. 1990.
[44] FHS, "Framingham Heart Study," 2018. [Online]. Available: https://www.framinghamheartstudy.org/about-the-fhs-participants/original-cohort/. [Accessed: 06-Dec-2018].
[45] J. Antony, K. Mcguinness, N. E. O'connor, and K. Moran, "Quantifying Radiographic Knee Osteoarthritis Severity using Deep Convolutional Neural Networks," 2016.
[46] M. Blagojevic, C. Jinks, A. Jeffery, and K. P. Jordan, "Risk factors for onset of osteoarthritis of the knee in older adults: a systematic review and meta-analysis," *Osteoarthr. Cartil.*, vol. 18, no. 1, pp. 24–33, 2010.
[47] G. Musumeci, F. C. Aiello, M. A. Szychlinska, M. Di Rosa, P. Castrogiovanni, and A. Mobasheri, "Osteoarthritis in the XXIst century: Risk factors and behaviours that influence disease onset and progression," *Int. J. Mol. Sci.*, vol. 16, no. 3, pp. 6093–6112, 2015.
[48] T. Felson, R. Lawrence, P. Dieppe, R. Hirsch, C. Helmick, and J. Jordan, "NIH Conference Osteoarthritis : New Insights. Part 1: The Disease and Its Risk Factors," *Ann. Intern. Med.*, vol. 133, no. 8, pp. 637–639, 2000.
[49] C. R. Chu, A. A. Williams, C. H. Coyle, and M. E. Bowers, "Early diagnosis to enable early treatment of pre-osteoarthritis," *Arthritis Res. Ther.*, vol. 14, no. 3, pp. 212–222, 2012.
[50] H. J. M. Kerkhof *et al.*, "Prediction model for knee osteoarthritis incidence, including clinical, genetic and biochemical risk factors," *Ann. Rheum. Dis.*, vol. 73, no. 12, pp. 2116–2121, 2014.
[51] D. Felson, J. Niu, D. Gross, and M. Englund, "Valgus Malaligment is a Risk Factor forLateral Knee Osteoarthritis Incidence and Progression: Findings from MOST and the Osteoarthritis Initiative," vol. 65, no. 2, pp. 355–362, 2014.
[52] W. F. T. Lai, C. H. Chang, Y. Tang, R. Bronson, and C. H. Tung, "Early diagnosis of osteoarthritis using cathepsin B sensitive near-infrared fluorescent probes," *Osteoarthr. Cartil.*, vol. 12, no. 3, pp. 239–244, 2004.
[53] W. Zhang *et al.*, "Nottingham knee osteoarthritis risk prediction models," *Ann. Rheum. Dis.*, vol. 70, no. 9, pp. 1599–1604, Sep. 2011.
[54] E. Losina, K. Klara, G. L. Michl, J. E. Collins, and J. N. Katz, "Development and feasibility of a personalized, interactive risk calculator for knee osteoarthritis," *BMC Musculoskelet. Disord.*, vol. 16, no. 1, p. 12, Oct. 2015.
[55] T. K. Yoo, D. W. Kim, S. B. Choi, E. Oh, J. Soo Park, and J. S. Park, "Simple Scoring System and Artificial Neural Network for Knee Osteoarthritis Risk Prediction: A Cross-Sectional Study," *PLoS One*, vol. 11, no. 2, p. e0148724 (17 pages), Feb. 2016.
[56] G. B. Joseph *et al.*, "Tool for osteoarthritis risk prediction (TOARP) over 8 years using baseline clinical data, X-ray, and MRI: Data from the osteoarthritis

initiative," *J. Magn. Reson. Imaging*, vol. 47, no. 6, pp. 1517–1526, Jun. 2018.
[57] Y. Wang *et al.*, "Causal Discovery in Radiographic Markers of Knee Osteoarthritis and Prediction for Knee Osteoarthritis Severity With Attention–Long Short-Term Memory," *Front. Public Heal.*, vol. 8, Dec. 2020.
[58] NHS, "BMI calculator | Check your BMI - NHS ," 2018. [Online]. Available: https://www.nhs.uk/live-well/healthy-weight/bmi-calculator/. [Accessed: 31-Aug-2021].
[59] "Diabetes UK – Know Your Risk of Type 2 diabetes." [Online]. Available: https://riskscore.diabetes.org.uk/start. [Accessed: 31-Aug-2021].
[60] "QRISK3." [Online]. Available: https://qrisk.org/three/index.php. [Accessed: 31-Aug-2021].
[61] M. KG *et al.*, "Risk prediction models: II. External validation, model updating, and impact assessment," *Heart*, vol. 98, no. 9, pp. 691–698, May 2012.
[62] NIMH, "OAI," *nda.nih.gov*. .
[63] M. C. Nevitt, D. T. Felson, and G. Lester, "OAI Protocol THE OSTEOARTHRITIS INITIATIVE PROTOCOL FOR THE COHORT STUDY," 2006.
[64] NDA, "OAI." [Online]. Available: https://nda.nih.gov/oai/study-details/schedule-of-assessments.html. [Accessed: 20-Oct-2019].
[65] D. T. Felson and M. C. Nevitt, "Epidemiologic studies for osteoarthritis: New versus conventional study design approaches," Rheum Dis Clin North Am, 2004.
[66] N. A. Segal *et al.*, "The Multicenter Osteoarthritis Study: Opportunities for Rehabilitation Research," *PM&R*, vol. 5, no. 8, pp. 647–654, Aug. 2013.
[67] "Osteoarthritis: Care and management in adults," 2014.
[68] N. Tamboli, "Tackling Missing Value in Dataset," *Analytics Vidhya*, 29-Oct-2021. [Online]. Available: https://www.analyticsvidhya.com/blog/2021/10/handling-missing-value/. [Accessed: 28-Mar-2022].
[69] K. Maladkar, "5 Ways To Handle Missing Values In Machine Learning Datasets," *Developers Corner*, 09-Feb-2018. [Online]. Available: https://analyticsindiamag.com/5-ways-handle-missing-values-machine-learning-datasets/. [Accessed: 28-Mar-2022].
[70] J. Barnard and X.-L. Meng, "Applications of multiple imputation in medical studies: from AIDS to NHANES," *Stat. Methods Med. Res.*, vol. 8, no. 1, pp. 17–36, Feb. 1999.
[71] H. Kang, "The prevention and handling of the missing data," *Korean J. Anesthesiol.*, vol. 64, no. 5, p. 402, May 2013.
[72] J. C. Jakobsen, C. Gluud, J. Wetterslev, and P. Winkel, "When and how should multiple imputation be used for handling missing data in randomised clinical trials - A practical guide with flowcharts," *BMC Med. Res. Methodol.*, vol. 17, no. 1, pp. 1–10, Dec. 2017.
[73] M. J. Azur, E. A. Stuart, C. Frangakis, and P. J. Leaf, "Multiple imputation by chained equations: what is it and how does it work?," *Int. J. Methods Psychiatr. Res.*, vol. 20, no. 1, p. 40, Mar. 2011.
[74] M. Jamshidian and M. Mata, "Advances in Analysis of Mean and Covariance Structure when Data are Incomplete," *Handb. Latent Var. Relat. Model.*, pp. 21–44, Jan. 2007.
[75] C. M. Salgado, C. Azevedo, H. Proença, and S. M. Vieira, "Missing Data," *Second. Anal. Electron. Heal. Rec.*, pp. 143–162, Jan. 2016.
[76] W. Zhang *et al.*, "Nottingham knee osteoarthritis risk prediction models," *Ann. Rheum. Dis.*, 2011.
[77] E. R. Vina and C. K. Kwoh, "Epidemiology of osteoarthritis: literature update,"

Curr. Opin. Rheumatol., vol. 30, no. 2, pp. 160–167, Mar. 2018.
[78] P. G. McCabe, I. Olier, S. Ortega-Martorell, I. Jarman, V. Baltzopoulos, and P. Lisboa, "Comparative Analysis for Computer-Based Decision Support: Case Study of Knee Osteoarthritis," in *Intelligent Data Engineering and Automated Learning -- IDEAL 2019*, 2019, pp. 114–122.
[79] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning with Applications in R Second Edition*, 2nd ed. Springer, 2021.
[80] E. W. Steyerberg and Y. Vergouwe, "Towards better clinical prediction models: seven steps for development and an ABCD for validation," *Eur. Heart J.*, vol. 35, no. 29, pp. 1925–1931, 2014.
[81] M. A. Ahmad, C. Eckert, A. Teredsai, and G. McKelvey, "Interpretable Machine Learning in Healthcare Muhammad Aurangzeb," *IEEE Intell. Informatics Bull.*, vol. 19, no. 1, pp. 1–7, 2018.
[82] B. Goodman and S. Flaxman, "European Union regulations on algorithmic decision-making and a 'right to explanation,'" 2016.
[83] G. C. M. Siontis, I. Tzoulaki, K. C. Siontis, and J. P. A. Ioannidis, "Comparisons of established risk prediction models for cardiovascular disease: systematic review," *BMJ Br. Med. J.*, vol. 344, p. e3318, May 2012.
[84] J. Kaijser, T. Bourne, B. Van Calster, D. Timmerman, and J. Kaijser, "Towards an evidence-based approach for diagnosis and management of adnexal masses: findings of the International Ovarian Tumour Analysis (IOTA) studies," *FaCTs Views Vis oBgyn*, vol. 7, no. 1, pp. 42–59, 2015.
[85] T. Lindeman, "3 Examples of How Hospitals are Using Predictive Analytics," *Dimensional Insight*, 2018. [Online]. Available: https://www.dimins.com/blog/2018/02/15/hospitals-predictive-analytics/. [Accessed: 29-Jul-2019].
[86] T. Therneau, B. Atkinson, and B. Ripley, "Package 'rpart.'" 10-Apr-2019.
[87] J. R. Quinlan, "Induction of Decision Trees," *Mach. Learn.*, pp. 81–106, 1985.
[88] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification and regression trees*, 1st ed. Boca Raton, 1984.
[89] M. E. Shipe, S. A. Deppen, F. Farjah, and E. L. Grogan, "Developing prediction models for clinical use using logistic regression: An overview," *J. Thorac. Dis.*, vol. 11, no. Suppl 4, pp. S574–S584, 2019.
[90] E. B. Ing and R. Ing, "The Use of a Nomogram to Visually Interpret a Logistic Regression Prediction Model for Giant Cell Arteritis," *Neuro-Ophthalmology*, vol. 42, no. 5, pp. 284–286, 2018.
[91] P. J. G. Lisboa, S. Ortega-Martorell, S. Cashman, and I. Olier, "The Partial Response Network: a neural network nomogram," Aug. 2019.
[92] P. J. Kindermans *et al.*, "Learning how to explain neural networks: PatternNet and PatternAttribution," *6th Int. Conf. Learn. Represent. ICLR 2018 - Conf. Track Proc.*, May 2017.
[93] B. Van Calster, D. Timmerman, I. T. Nabney, L. Valentin, C. Van Holsbeke, and S. Van Huffel, "Classifying ovarian tumors using Bayesian multi-layer perceptrons and automatic relevance determination: A multi-center study," *Annu. Int. Conf. IEEE Eng. Med. Biol. - Proc.*, pp. 5342–5345, 2006.
[94] D. Husmeier, "Automatic Relevance Determination (ARD)," in *Neural Networks for Conditional Probability Estimation: Forecasting Beyond Point Predictions*, London: Springer London, 1999, pp. 221–227.
[95] A. Blumer, A. Ehrenfeucht, D. Haussler, and M. K. Warmuth, "Occam's Razor," *Inf. Process. Lett.*, vol. 24, no. 6, 1987.
[96] D. J. C. MacKay, *Models of Neural Networks III*. New York, NY: Springer New

York, 1996.
[97] R. Mbuvha, I. Boulkaibet, and T. Marwala, "Automatic Relevance Determination Bayesian Neural Networks for Credit Card Default Modelling."
[98] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognit. Lett.*, vol. 27, no. 8, pp. 861–874, 2006.
[99] A. Iasonos, D. Schrag, G. V. Raj, and K. S. Panageas, "How to build and interpret a nomogram for cancer prognosis," *J. Clin. Oncol.*, vol. 26, no. 8, pp. 1364–1370, Mar. 2008.
[100] Q. Li, K. Amano, T. M. Link, and C. B. Ma, "Advanced Imaging in Osteoarthritis," *Sports Health*, vol. 8, no. 5, p. 418, Sep. 2016.
[101] S. Curtis, "Osteoarthritis Diagnosis," *Arthritis Health*, 2020. [Online]. Available: https://www.arthritis-health.com/types/osteoarthritis/osteoarthritis-diagnosis. [Accessed: 18-Mar-2022].
[102] Q. Li, K. Amano, T. M. Link, and C. B. Ma, "Advanced Imaging in Osteoarthritis," *Sports Health*, vol. 8, no. 5, p. 418, Sep. 2016.
[103] "How fast does osteoarthritis progress? Does it evolve steadily?," *Arthrolink*, 2020. [Online]. Available: https://www.arthrolink.com/en/disease/knowing/evolution-osteoarthritis. [Accessed: 17-Jun-2020].
[104] L. L. Johnson, *An Introduction to Survival Analysis*. Elsevier Inc., 2018.
[105] T. Zahid, "Survival Analysis ," *TowardsDataScience*, 18-Mar-2019. [Online]. Available: https://towardsdatascience.com/survival-analysis-part-a-70213df21c2e. [Accessed: 05-Feb-2021].
[106] R. L. Prentice and J. D. Kalbfleisch, "Survival Analysis: Overview," *Int. Encycl. Soc. Behav. Sci.*, pp. 15318–15325, Jan. 2001.
[107] V. S. Stel, F. W. Dekker, G. Tripepi, C. Zoccali, and K. J. Jager, "Survival Analysis II: Cox Regression," *Kidney Dis. Popul. Heal.* , vol. 119, pp. 255–260, 2011.
[108] V. S. Stel, F. W. Dekker, G. Tripepi, C. Zoccali, and K. J. Jager, "Survival Analysis I: The Kaplan-Meier Method," *Kidney Dis. Popul. Heal.*, vol. 119, pp. 83–88, 2011.
[109] C. T. C. Arsene and P. J. G. Lisboa, "Artificial Neural Networks Used in the Survival Analysis of Breast Cancer Patients: A Node-Negative Study," *Outcome Predict. Cancer*, pp. 191–239, Jan. 2007.
[110] R. H. Sprague and H. J. Watson, "Decision support systems: Putting theory into practice," 1986.
[111] W. N. Dudley, PhD, R. Wickham, PhD, RN, AOCN, and N. Coombs, MS, "An Introduction to Survival Statistics: Kaplan-Meier Analysis," *J. Adv. Pract. Oncol.*, vol. 7, no. 1, p. 91, Feb. 2016.
[112] M. Kudo *et al.*, "Survival Analysis over 28 Years of 173,378 HCC Patients in," *Japan Liver Cancer*, vol. 5, pp. 190–197, 2016.
[113] M. Luck, T. Sylvain, A. Lodi, and Y. Bengio, "Deep Learning for Patient-Specific Kidney Graft Survival Analysis," 2017.
[114] S. Annibali, N. Pranno, M. P. Cristalli, G. La Monaca, and A. Polimeni, "Survival Analysis of Implant in Patients with Diabetes Mellitus: A Systematic Review," *Implant Dentistry*, vol. 25, no. 5. Lippincott Williams and Wilkins, pp. 663–674, 01-Oct-2016.
[115] J. S. Everhart, M. M. Abouljoud, J. Kirven, and D. C. Flanigan, "Full-Thickness Cartilage Defects Are Important Independent Predictive Factors for Progression to Total Knee Arthroplasty in Older Adults with Minimal to Moderate Osteoarthritis," *J. Bone Jt. Surg.*, vol. 101, no. 1, pp. 56–63, 2019.

[116] K. Leung *et al.*, "Prediction of total knee replacement and diagnosis of osteoarthritis by using deep learning on knee radiographs: Data from the osteoarthritis initiative," *Radiology*, vol. 296, no. 3, pp. 584–593, Sep. 2020.
[117] S. R. W. Wijn, M. M. Rovers, T. G. Van Tienen, and G. Hannink, "Intra-articular corticosteroid injections increase the risk of requiring knee arthroplasty a multicentre longitudinal observational study using data from the osteoarthritis initiative," *Bone Jt. J.*, vol. 102-B, no. 5, pp. 586–592, May 2020.
[118] C. Zeng *et al.*, "Intra-articular corticosteroids and the risk of knee osteoarthritis progression: results from the Osteoarthritis Initiative," *Osteoarthr. Cartil.*, vol. 27, no. 6, pp. 855–862, Jun. 2019.
[119] F. Wolfe and N. E. Lane, "The longterm outcome of osteoarthritis: rates and predictors of joint space narrowing in symptomatic patients with knee osteoarthritis.," *J. Rheumatol.*, vol. 29, no. 1, 2002.
[120] V. Vennu, H. Misra, and A. Misra, "Depressive symptoms and the risk of arthritis: A survival analysis using data from the osteoarthritis initiative," *Indian J. Psychiatry*, vol. 61, no. 5, pp. 444–450, Sep. 2019.
[121] S. Törmälehto, E. Aarnio, M. E. Mononen, J. P. A. Arokoski, R. K. Korhonen, and J. A. Martikainen, "Eight-year trajectories of changes in health-related quality of life in knee osteoarthritis: Data from the Osteoarthritis Initiative (OAI)," *PLoS One*, vol. 14, no. 7, p. e0219902, Jul. 2019.
[122] "What is the body mass index (BMI)? - NHS," 15-Jul-2019. [Online]. Available: https://www.nhs.uk/common-health-questions/lifestyle/what-is-the-body-mass-index-bmi/. [Accessed: 29-Jul-2020].
[123] İ. Etikan, K. Bukirova, and M. Yuvalı, "Choosing statistical tests for survival analysis," 2018.
[124] D. E. Matthews and V. T. Farewell, *Using and Understanding Medical Statistics*. 2015.
[125] D. G. Kleinbaum, "The Cox Proportional Hazards Model and Its Characteristics," in *Survival Analysis: A Self-Learning Text*, New York, NY: Springer New York, 1996, pp. 83–128.
[126] Y. Yao, "Several Methods to assess proportional hazard assumption when applying COX regression model," Shanghai, 2018.
[127] M. J. Bradburn, T. G. Clark, S. B. Love, and D. G. Altman, "Survival Analysis Part II: Multivariate data analysis-an introduction to concepts and methods," *Br. J. Cancer*, vol. 89, pp. 431–436, 2003.
[128] R. Singh and K. Mukhopadhyay, "Survival analysis in clinical trials: Basics and must know areas," *Perspect. Clin. Res.*, vol. 2, no. 4, p. 145, 2011.
[129] D. Schutte, "Survival Analysis in R For Beginners - DataCamp," *DataCamp*, 2019. [Online]. Available: https://www.datacamp.com/community/tutorials/survival-analysis-R. [Accessed: 09-Jul-2020].
[130] B. George, S. Seals, and I. Aban, "Survival analysis and regression models," *Journal of Nuclear Cardiology*, vol. 21, no. 4. Springer New York LLC, pp. 686–694, 2014.
[131] H. Akaike, "A New Look at the Statistical Model Identification," *IEEE Trans. Automat. Contr.*, vol. 19, no. 6, pp. 716–723, 1974.
[132] D. Schoenfeld, "Partial residuals for the proportional hazards regression model," *Biometrika*, vol. 69, no. 1, pp. 239–280, 1982.
[133] M. Sekhon, M. Cartwright, and J. J. Francis, "Acceptability of healthcare interventions: an overview of reviews and development of a theoretical framework," *BMC Heal. Serv. Res. 2017 171*, vol. 17, no. 1, pp. 1–13, Jan. 2017.
[134] T. Martin and A. Poor, "How Your Body Heals After You Quit Smoking,"

VeryWellMind, 03-Apr-2020. [Online]. Available: https://www.verywellmind.com/after-the-last-cigarette-how-your-body-heals-2824388. [Accessed: 03-Feb-2021].
[135] E. D. Gometz, "CLINICAL PEARL Health Effects of Smoking and the Benefits of Quitting," 2011.
[136] F. E. Harrell, R. M. Califf, D. B. Pryor, K. L. Lee, and R. A. Rosati, "Evaluating the Yield of Medical Tests," *JAMA J. Am. Med. Assoc.*, vol. 247, no. 18, pp. 2543–2546, May 1982.
[137] J. Barata, "What is Model Validation?," *Yields.io*, 03-Feb-2020. [Online]. Available: https://www.yields.io/blog/what-is-model-validation/. [Accessed: 13-Jul-2021].
[138] A. E. Ivanescu *et al.*, "The importance of prediction model validation and assessment in obesity and nutrition research," *Int. J. Obes. 2016 406*, vol. 40, no. 6, pp. 887–894, Oct. 2015.
[139] T. L. Paez, "Introduction to Model Validation," Albuquerque, 2009.
[140] C. A. Aumann, "A methodology for developing simulation models of complex systems," *Ecol. Modell.*, vol. 202, no. 3–4, pp. 385–396, Apr. 2007.
[141] K. Krishnan and M. Andersen, "Physiologically Based Pharmacokinetic Models in the Risk Assessment of Developmental Neurotoxicants," in *Handbook of Developmental Neurotoxicology*, Academic Press, 1998, pp. 709–725.
[142] D. A. Bluemke *et al.*, "Assessing Radiology Research on Artificial Intelligence: A Brief Guide for Authors, Reviewers, and Readers—From the Radiology Editorial Board," *https://doi.org/10.1148/radiol.2019192515*, vol. 294, no. 2, pp. 487–489, Dec. 2019.
[143] F. E. Harrell, *Regression Modeling Strategies*, Second. Cham: Springer International Publishing, 2015.
[144] T. DB, J. KJ, V. Y, and M. KG, "Validation, updating and impact of clinical prediction rules: a review," *J. Clin. Epidemiol.*, vol. 61, no. 11, pp. 1085–1094, Nov. 2008.
[145] J. Tohka and M. van Gils, "Evaluation of machine learning algorithms for health and wellness applications: A tutorial," *Comput. Biol. Med.*, vol. 132, p. 104324, May 2021.
[146] L. Wang, H. Lu, H. Chen, S. Jin, M. Wang, and S. Shang, "Development of a model for predicting the 4-year risk of symptomatic knee osteoarthritis in China: a longitudinal cohort study," *Arthritis Res. Ther.*, vol. 23, no. 1, 2021.
[147] B. Sheng *et al.*, "Identification of knee osteoarthritis based on bayesian network: Pilot study," *JMIR Med. Informatics*, vol. 7, no. 3, 2019.
[148] K. Magnusson, A. Turkiewicz, S. Timpka, and M. Englund, "A prediction model for the 40-year risk of knee osteoarthritis in adolescent men," *Arthritis Care Res.*, vol. 71, no. 4, pp. 558–562, 2019.
[149] A. Tiulpin, J. Thevenot, E. Rahtu, P. Lehenkari, and S. Saarakkala, "Automatic knee osteoarthritis diagnosis from plain radiographs: A deep learning-based approach," *Sci. Rep.*, vol. 8, no. 1, pp. 1–10, Dec. 2018.
[150] G. S. Fernandes, A. Bhattacharya, D. F. McWilliams, S. L. Ingham, M. Doherty, and W. Zhang, "Risk prediction model for knee pain in the Nottingham community: A Bayesian modelling approach," *Arthritis Res. Ther.*, vol. 19, no. 1, pp. 1–8, 2017.
[151] P. Widera *et al.*, "Multi-classifier prediction of knee osteoarthritis progression from incomplete imbalanced longitudinal data," *Sci. Rep.*, vol. 10, no. 1, pp. 1–15, 2020.
[152] "NHS England » NHS to rollout lung cancer scanning trucks across the

country," 08-Feb-2019. [Online]. Available: https://www.england.nhs.uk/2019/02/lung-trucks/. [Accessed: 30-Aug-2021].
[153] T. D. Pigott, "A Review of Methods for Missing Data," 2001.
[154] T. M. Therneau, T. Lumley, E. Atkinson, and C. Crowson, "Package 'survival' ," 2021.
[155] S. Potapov, W. Adler, and M. Schmid, "Package 'survAUC' Estimators of prediction accuracy for time-to-event data," Feb. 2015.
[156] A. Kassambara, M. Kosinski, B. Przemyslaw, and F. Scheipl, "Package 'survminer' ," 2021.
[157] "Package 'shiny,'" 2021.
[158] X. Robin *et al.*, "Package 'pROC,'" Jan. 2021.
[159] M. Kuhn *et al.*, "Package 'caret' ," Mar. 2020.
[160] S. Zashin and C. Eustice, "Symptomatic vs. Radiographic Osteoarthritis," *Verywell Health*, 17-Oct-2020. [Online]. Available: https://www.verywellhealth.com/symptomatic-osteoarthritis-radiographic-osteoarthritis-2552211. [Accessed: 17-Dec-2020].
[161] T. J. Bright *et al.*, "Effect of Clinical Decision-Support Systems," *Ann. Intern. Med.*, vol. 157, no. 1, p. 29, Jul. 2012.
[162] P. M. Ravdin *et al.*, "Computer program to assist in making decisions about adjuvant therapy for women with early breast cancer," *J. Clin. Oncol.*, vol. 19, no. 4, pp. 980–991, Feb. 2001.
[163] D. Timmerman *et al.*, "Simple ultrasound-based rules for the diagnosis of ovarian cancer," *Ultrasound Obstet. Gynecol.*, vol. 31, no. 6, pp. 681–690, Jun. 2008.
[164] D. Timmerman *et al.*, "Predicting the risk of malignancy in adnexal masses based on the Simple Rules from the International Ovarian Tumor Analysis group," *Am. J. Obstet. Gynecol.*, pp. 424–437, 2016.
[165] T. K. Yoo, D. W. Kim, S. B. Choi, E. Oh, and J. Soo Park, "Simple Scoring System and Artificial Neural Network for Knee Osteoarthritis Risk Prediction: A Cross-Sectional Study," *PLoS One*, vol. 11, no. 2, 2016.
[166] N. D. Smith, "Plato and Aristotle on the Nature of Women," *J. Hist. Philos.*, vol. 21, no. 4, pp. 467–478, 1983.
[167] P. C. Elwood *et al.*, "A Randomized Controlled Trial of Acetyl Salicyclic Acid in the Secondary Prevention of Mortality from Myocardial Infarction," *Br. Med. J.*, vol. 1, no. 5905, p. 436, Mar. 1974.
[168] J. R. O'brien *et al.*, "Regular Aspirin Intake and Acute Myocardial Infarction," *Br. Med. J.*, vol. 1, no. 5905, p. 440, Mar. 1974.
[169] L. Schiebinger, "Women's health and clinical trials," *J. Clin. Invest.*, vol. 112, no. 7, p. 973, Oct. 2003.
[170] I. of Medicine, *Exploring the Biological Contributions to Human Health: Does Sex Matter?* Washington, DC: The National Academies Press, 2001.
[171] K. A. Liu and N. A. D. Mager, "Women's involvement in clinical trials: historical perspective and future implications," *Pharm. Pract. (Granada).*, vol. 14, no. 1, 2016.
[172] B. A. Mikulski, *Women's Health Equity Act of 1991*. 102D Congress, 1991.
[173] *Women's Health Equity Act of 1993*. 103D Congress, 1993.
[174] J. Billock, "Pain bias: The health inequality rarely discussed," *BBC*, 22-May-2018. [Online]. Available: https://www.bbc.com/future/article/20180518-the-inequality-in-how-women-are-treated-for-pain. [Accessed: 06-Sep-2021].
[175] M. Dusenbery, "'Everybody was telling me there was nothing wrong' ," *BBC*, 29-May-2018. [Online]. Available: https://www.bbc.com/future/article/20180523-how-gender-bias-affects-your-healthcare. [Accessed: 06-Sep-2021].
[176] Mayo Clinic Staff, "Depression in women: Understanding the gender gap," 2019.

[Online]. Available: https://www.mayoclinic.org/diseases-conditions/depression/in-depth/depression/art-20047725. [Accessed: 06-Sep-2021].
[177] A. Lyons, "The 'stoic' stereotype and its impact on men's mental health," *RACGP*, 2018. [Online]. Available: https://www1.racgp.org.au/newsgp/professional/the-'stoic'-stereotype-and-its-impact-on-men's-men. [Accessed: 06-Sep-2021].
[178] B. Floyd, "Problems in accurate medical diagnosis of depression in female patients.," *Soc. Sci. Med.*, vol. 44, no. 3, pp. 403–412, Feb. 1997.
[179] R. Y. Khamis, T. Ammari, and G. W. Mikhail, "Gender differences in coronary heart disease," *Heart*, vol. 102, no. 14, pp. 1142–1149, Jul. 2016.
[180] S. K. Keitt, T. F. Fagan, and S. A. Marts, "Understanding sex differences in environmental health: a thought leaders' roundtable.," *Environ. Health Perspect.*, vol. 112, no. 5, p. 604, 2004.
[181] N. A. Karp *et al.*, "Prevalence of sexual dimorphism in mammalian phenotypic traits," *Nat. Commun. 2017 81*, vol. 8, no. 1, pp. 1–12, Jun. 2017.
[182] M. L. Blair, "Sex-based differences in physiology: what should we teach in the medical curriculum?," *https://doi.org/10.1152/advan.00118.2006*, vol. 31, no. 1, pp. 23–25, 2007.
[183] M. W. Lear, "Opinion | The Woman's Heart Attack," *The New York Times*, 26-Sep-2014.
[184] G. Belz and C. Seillet, "Man flu is real, but women get more autoimmune diseases and allergies," 2017. [Online]. Available: https://theconversation.com/man-flu-is-real-but-women-get-more-autoimmune-diseases-and-allergies-77248. [Accessed: 06-Sep-2021].
[185] M. Cimons, "Why do autoimmune diseases affect women more often than men? - The Washington Post," 2016. [Online]. Available: https://www.washingtonpost.com/national/health-science/why-do-autoimmune-diseases-affect-women-more-often-than-men/2016/10/17/3e224db2-8429-11e6-ac72-a29979381495_story.html. [Accessed: 06-Sep-2021].
[186] S. Reardon, "Infections reveal inequality between the sexes | Nature," *Nature*, 2016. [Online]. Available: https://www.nature.com/articles/534447a. [Accessed: 06-Sep-2021].
[187] S. Reardon, "Infections reveal inequality between the sexes," *Nature*, vol. 534, no. 7608, p. 447, Jun. 2016.
[188] S. L. Klein and A. Pekosz, "Sex-based Biology and the Rational Design of Influenza Vaccination Strategies," *J. Infect. Dis.*, vol. 209, no. Suppl 3, p. S114, 2014.
[189] M. Berg, Y. Appelman, and M. Bekker, "Gender and Health Knowledge Agenda," 2015.
[190] M. I. O'Connor, "Sex differences in osteoarthritis of the hip and knee," *J. Am. Acad. Orthop. Surg.*, vol. 15, no. SUPPL. 1, 2007.
[191] M. H. Laitner, L. C. Erickson, and E. Ortman, "Understanding the Impact of Sex and Gender in Osteoarthritis: Assessing Research Gaps and Unmet Needs," *J. Women's Heal.*, vol. 30, no. 5, pp. 634–641, May 2021.
[192] D. Prieto-Alhambra, A. Judge, M. K. Javaid, C. Cooper, A. Diez-Perez, and N. K. Arden, "'Incidence and risk factors for clinically diagnosed knee, hip and hand osteoarthritis: influences of age, gender and osteoarthritis affecting other joints,'" *Ann. Rheum. Dis.*, vol. 73, no. 9, p. 1659, 2014.
[193] E. Losina *et al.*, "Lifetime risk and age of diagnosis of symptomatic knee

osteoarthritis in the US," *Arthritis Care Res. (Hoboken).*, vol. 65, no. 5, pp. 703–711, May 2013.
[194] V. Srikanth, J. Fryer, G. Zhai, T. Winzenberg, D. Hosmer, and G. Jones, "A meta-analysis of sex differences prevalence, incidence and severity of osteoarthritis," *Osteoarthr. Cartil.*, vol. 13, no. 9, pp. 769–781, Sep. 2005.
[195] S. L. Hame and R. A. Alexander, "Knee osteoarthritis in women," *Curr. Rev. Musculoskelet. Med.*, vol. 6, no. 2, p. 182, Jun. 2013.
[196] D. Felson, A. Naimark, J. Anderson, L. Kazis, W. Castelli, and R. Meenan, "The prevalence of knee osteoarthritis in the elderly. The Framingham Osteoarthritis Study," *Arthritis Rheum.*, vol. 30, no. 8, pp. 914–918, 1987.
[197] A. Bracilovic, "Why Are Women More Prone to Osteoarthritis?," *Arthritis-health.com*, 2021. [Online]. Available: https://www.arthritis-health.com/blog/why-are-women-more-prone-osteoarthritis. [Accessed: 03-Sep-2021].
[198] F. Hanna *et al.*, "Women have increased rates of cartilage loss and progression of cartilage defects at the knee than men: a gender study of adults without clinical knee osteoarthritis," *Menopause*, vol. 16, no. 4, pp. 666–670, Jul. 2009.
[199] B. Heidari, "Knee osteoarthritis prevalence, risk factors, pathogenesis and features: Part I," *Casp. J. Intern. Med.*, vol. 2, no. 2, p. 205, Mar. 2011.
[200] J. Chappell, B. Yu, D. Kirkendall, and W. Garrett, "A comparison of knee kinetics between male and female recreational athletes in stop-jump tasks," *Am. J. Sports Med.*, vol. 30, no. 2, pp. 261–267, 2002.
[201] K. Ford, G. Myer, and T. Hewett, "Valgus knee motion during landing in high school female and male basketball players," *Med. Sci. Sports Exerc.*, vol. 35, no. 10, pp. 1745–1750, Oct. 2003.
[202] K. Ford, G. Myer, H. Toms, and T. Hewett, "Gender differences in the kinematics of unanticipated cutting in young athletes," *Med. Sci. Sports Exerc.*, vol. 37, no. 1, pp. 124–129, Jan. 2005.
[203] J. Mendiguchia, K. R. Ford, C. E. Quatman, E. Alentorn-Geli, and T. E. Hewett, "Sex Differences in Proximal Control of the Knee Joint," *Sports Med.*, vol. 41, no. 7, p. 541, 2011.
[204] T. W. Kernozek, M. R. Torry, H. Van Hoof, H. Cowley, and S. Tanner, "Gender Differences in Frontal and Sagittal Plane Biomechanics during Drop Landings," *Med. Sci. Sport. Exerc*, vol. 37, no. 6, pp. 1003–1012, 2005.
[205] F. Cicuttini, A. Forbes, K. Morris, S. Darling, M. Bailey, and S. Stuckey, "Gender differences in knee cartilage volume as measured by magnetic resonance imaging," *Osteoarthr. Cartil.*, vol. 7, pp. 265–271, 1999.
[206] S. Tummala, D. Schiphof, I. Byrjalsen, and E. B. Dam, "Gender Differences in Knee Joint Congruity Quantified from MRI: A Validation Study with Data from Center for Clinical and Basic Research and Osteoarthritis Initiative:," *https://doi.org/10.1177/1947603516684590*, vol. 9, no. 1, pp. 38–45, Dec. 2016.
[207] C. Chu *et al.*, "The feasibility of randomized controlled trials for early arthritis therapies (Earth) involving acute anterior cruciate ligament tear cohorts," *Am. J. Sports Med.*, vol. 40, no. 11, pp. 2648–2652, Nov. 2012.
[208] F. Nelson *et al.*, "Early post-traumatic osteoarthritis-like changes in human articular cartilage following rupture of the anterior cruciate ligament," *Osteoarthr. Cartil.*, vol. 14, no. 2, pp. 114–119, Feb. 2006.
[209] N. Friel and C. Chu, "The role of ACL injury in the development of posttraumatic knee osteoarthritis," *Clin. Sports Med.*, vol. 32, no. 1, pp. 1–12, Jan. 2013.
[210] X. Jin *et al.*, "Associations between endogenous sex hormones and MRI

structural changes in patients with symptomatic knee osteoarthritis," *Osteoarthr. Cartil.*, vol. 25, no. 7, pp. 1100–1106, Jul. 2017.
[211] R. S. Richmond, C. S. Carlson, T. C. Register, G. Shanker, and R. F. Loeser, "FUNCTIONAL ESTROGEN RECEPTORS IN ADULT ARTICULAR CARTILAGE Estrogen Replacement Therapy Increases Chondrocyte Synthesis of Proteoglycans and Insulin-Like Growth Factor Binding Protein 2," *ARTHRITIS Rheum.*, vol. 43, no. 9, pp. 2081–2090, 2000.
[212] M. I. O'Connor and E. G. Hooten, "Breakout Session: Gender Disparities in Knee Osteoarthritis and TKA," *Clin. Orthop. Relat. Res.*, vol. 469, no. 7, p. 1883, 2011.
[213] N. D. Clement, D. Weir, J. Holland, and D. J. Deehan, "Sex does not clinically influence the functional outcome of total knee arthroplasty but females have a lower rate of satisfaction with pain relief," *Knee Surg. Relat. Res. 2020 321*, vol. 32, no. 1, pp. 1–7, Jun. 2020.
[214] M. A. Ritter, J. T. Wing, M. E. Berend, K. E. Davis, and J. B. Meding, "The Clinical Effect of Gender on Outcome of Total Knee Arthroplasty," *J. Arthroplasty*, vol. 23, no. 3, pp. 331–336, Apr. 2008.
[215] D. F. Dalury, J. B. Mason, J. A. Murphy, and M. J. Adams, "Analysis of the outcome in male and female patients using a unisex total knee replacement system," *J. Bone Jt. Surg. - Ser. B*, vol. 91, no. 3, pp. 357–360, Mar. 2009.
[216] B. L. Wise *et al.*, "The association of parity with osteoarthritis and knee replacement in the Multicenter Osteoarthritis Study," *Osteoarthr. Cartil.*, vol. 21, no. 12, pp. 1849–1854, Dec. 2013.
[217] B. Liu, A. Balkwill, C. Cooper, A. Roddam, A. Brown, and V. Beral, "Reproductive history, hormonal factors and the incidence of hip and knee replacement for osteoarthritis in middle-aged women," *Ann. Rheum. Dis.*, vol. 68, no. 7, pp. 1165–1170, Jul. 2009.
[218] B. M. de Klerk *et al.*, "Limited evidence for a protective effect of unopposed oestrogen therapy for osteoarthritis of the hip: a systematic review," *Rheumatology*, vol. 48, no. 2, pp. 104–112, Feb. 2009.
[219] S. C. Suddarth and Y. L. Kergosien, *Rule-injection hints as a means of improving network performance and learning time*, vol. 412 LNCS. 1990.
[220] Y. S. Abu-Mostafa, "Learning from hints in neural networks," *J. Complex.*, vol. 6, no. 2, pp. 192–198, Jun. 1990.
[221] C. Ciliberto, Y. Mroueh, T. Poggio, and L. Rosasco, "Convex Learning of Multiple Tasks and their Structure," pp. 1–26, 2015.
[222] P. Jawanpuria and J. S. Nath, "A convex feature learning formulation for latent task structure discovery," *Proc. 29th Int. Conf. Mach. Learn. ICML 2012*, vol. 1, pp. 137–144, 2012.
[223] A. Kumar and H. Daumé III, "Learning Task Grouping and Overlap in Multi-Task Learning," 2010.
[224] S. Zhong, J. Pu, Y.-G. Jiang, R. Feng, and X. Xue, "Flexible multi-task learning with latent task grouping," 2015.
[225] Caruana and R. M. Seraj, "Multi-task Learning," *Mach. Learn.*, vol. 28, no. 1, pp. 41–75, 1997.
[226] B. Romera-Paredes, A. Argyriou, N. Bianchi-Berthouze, M. Pontil, N. Berthouze, and M. Pontil, "Exploiting Unrelated Tasks in Multi-Task Learning," *Proc. 15th Int. Conf. Artif. Intell. Stat.*, vol. 22, pp. 951–959, 2012.
[227] A. Evgeniou and M. Pontil, "Multi-task feature learning," pp. 243–272, 2007.
[228] C. Szegedy *et al.*, "Going deeper with convolutions," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 07-12-June, pp. 1–9, 2015.

[229] G. Roig, "Learning Data Representation : DNN Tips and Tricks," 2015.
[230] K. Murugesan and J. Carbonell, "Self-Paced Multitask Learning with Shared Knowledge," 2017.
[231] A. Zweig and G. Chechik, "Group online adaptive learning," *Mach. Learn.*, vol. 106, no. 9–10, pp. 1747–1770, 2017.
[232] M. R. Masliah, "Stationarity/Nonstationarity Identification." [Online]. Available: http://etclab.mie.utoronto.ca/people/moman/Stationarity/stationarity.html. [Accessed: 20-Aug-2018].
[233] K. Murugesan, L.-P. Morency, and C. Barnabás Póczos, "Online and Adaptive Methods for Multitask Learning," 2017.
[234] R. S. Simões, V. G. Maltarollo, P. R. Oliveira, and K. M. Honorio, "Transfer and Multi-task Learning in QSAR Modeling: Advances and Challenges.," *Front. Pharmacol.*, vol. 9, p. 74, 2018.
[235] W. Dai, Q. Yang, G.-R. Xue, and Y. Yu, "Boosting for transfer learning," in *Proceedings of the 24th international conference on Machine learning - ICML '07*, 2007, pp. 193–200.
[236] R. Raina, A. Battle, H. Lee, B. Packer, and A. Y. Ng, "Self-taught learning," in *Proceedings of the 24th international conference on Machine learning - ICML '07*, 2007, pp. 759–766.
[237] N. D. Lawrence and J. C. Platt, "Learning to learn with the informative vector machine," in *Twenty-first international conference on Machine learning - ICML '04*, 2004, p. 65.
[238] L. Mihalkova, T. Huynh, and R. J. Mooney, "Mapping and Revising Markov Logic Networks for Transfer Learning," 2007.
[239] S. Ruder, "An Overview of Multi-Task Learning in Deep Neural Networks," no. May, 2017.
[240] Y. Zhang and Q. Yang, "A Survey on Multi-Task Learning," *IEEE Trans. Knowl. Data Eng.*, 2021.
[241] M. Crawshaw, "MULTI-TASK LEARNING WITH DEEP NEURAL NETWORKS: A SURVEY," 2020.
[242] A. de la V. de León, B. Chen, and V. J. Gillet, "Effect of missing data on multitask prediction methods," *J. Cheminform.*, vol. 10, no. 1, p. 26, Dec. 2018.
[243] H. Yuan, I. Paskov, H. Paskov, A. J. González, and C. S. Leslie, "Multitask learning improves prediction of cancer drug sensitivity," *Sci. Rep.*, vol. 6, no. 1, p. 31619, Oct. 2016.
[244] J. Attenberg, K. Weinberger, A. Dasgupta, A. Smola, and M. Zinkevich, "Collaborative Email-Spam Filtering with the Hashing-Trick," 2009.
[245] O. Chapelle, P. Shivaswamy, S. Vadrevu, K. Weinberger, Y. Zhang, and B. Tseng, "Multi-Task Learning for Boosting with Application to Web Search Ranking," in *KDD '10: Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2010.
[246] S. Roy *et al.*, "Multitask prediction of organ dysfunction in the intensive care unit using sequential subnetwork routing," *J. Am. Med. Informatics Assoc.*, vol. 28, no. 9, pp. 1936–1946, Aug. 2021.
[247] S. Thrun, "Is Learning The n-th Thing Any Easier Than Learning The First?"
[248] O. Chapelle, P. Shivaswamy, S. Vadrevu, K. Weinberger, Y. Zhang, and B. Tseng, "Boosted multi-task learning," *Mach. Learn.*, vol. 85, no. 1–2, pp. 149–173, Oct. 2011.
[249] N. Srivastava, G. Hinton, A. Krizhevsky, and R. Salakhutdinov, "Dropout: A Simple Way to Prevent Neural Networks from Overfitting," *J. Mach. Learn. Res.*, vol. 15, pp. 1929–1958, 2014.

[250] P. Ramachandran, B. Zoph, and Q. V Le Google Brain, "SEARCHING FOR ACTIVATION FUNCTIONS," p. 13, Oct. 2017.
[251] J. He, L. Li, J. Xu, and C. Zheng, "ReLU Deep Neural Networks and Linear Finite Elements," 2018.
[252] A. Goldstein, A. Kapelner, J. Bleich, and E. Pitkin, "Peeking Inside the Black Box: Visualizing Statistical Learning with Plots of Individual Conditional Expectation," 2014.
[253] E. M. Roos and L. S. Lohmander, "Health and Quality of Life Outcomes The Knee injury and Osteoarthritis Outcome Score (KOOS): from joint injury to osteoarthritis The Knee injury and Osteoarthritis Outcome Score (KOOS)," 2003.
[254] V. Silverwood, M. Blagojevic-Bucknall, C. Jinks, J. L. Jordan, J. Protheroe, and K. P. Jordan, "Current evidence on risk factors for knee osteoarthritis in older adults: a systematic review and meta-analysis," *Osteoarthr. Cartil.*, vol. 23, no. 4, pp. 507–515, Apr. 2015.
[255] M. Vogel *et al.*, "Validation of Myocardial Acceleration During Isovolumic Contraction as a Novel Noninvasive Index of Right Ventricular Contractility," *Circulation*, vol. 105, no. 14, pp. 1693–1699, Apr. 2002.
[256] N. Barda *et al.*, "Developing a COVID-19 mortality risk prediction model when individual-level data are not available."

Glossary

Abbreviation	Definition
ACL	Anterior cruciate ligament
AIC	Akaike information criterion
ANN	Artificial neural network
ARD	Automatic relevance determination
AUC	Area under the curve
AUROC	Area under the receiver operating characteristic curve
CADx	Computer aided diagnosis
CART	Classification and regression trees
CHAID	Chi-squared automatic interaction detection
CI	Confidence interval
CNN	Convolutional neural network
GDP	Gross domestic product
GDPR	General data protection regulations
HMO	Health membership organisation
HR	Hazard ratio
KL	Kellgren-Lawrence
KM	Kaplan-Meier
KNHANES	Korean National Health and Nutrition Examination Survey
KOA	Knee osteoarthritis
Lasso	Least absolute shrinkage and selection operator
LIME	Locally interpretable model agnostic explanation
LogR/LR	Logistic regression
MICE	Multiple imputation by chained equations
ML	Machine learning
MLP	Multilayer perceptron
MLP-ARD	Multilayer perceptron Automatic relevance determination
MOST	Multicentre Osteoarthritis Study
MTL	Multitask learning
NHS	National Health Service
NIRF	Near infrared fluorescence
NN	Neural Network
NSAIDs	Nonsteroidal anti-inflammatory drugs
OA	Osteoarthritis
OAI	Osteoarthritis Initiative
OSRE	Orthogonal search rule extraction
PPV	Positive predictive value
PRN	Partial response network
ReLu	Rectified linear unit
RF	Random forest

ROC	Receiver operating characteristic
STL	Single task learning
SVM	Support vector machine

Appendices

In Chapter 2, there is a summary of the datasets used throughout this thesis. Provided here are tables showing the data characteristics for the OAI, MOST and OActive data used for any diagnostic modelling mentioned in the thesis.

OAI

The OAI data consists of 1187 variables collecting information relating to family history, medical history, medication, physical activity, pain and symptoms, along with demographic information. Table 0-1 contains a summary of information relating to variable subset used for the diagnostic modelling.

Table 0-1: Summary table for the OAI data detailing the variables used within the thesis.

	0 (N=2473)	**1 (N=1997)**	**Total (N=4470)**
AGE			
Mean (SD)	59.850 (9.130)	62.647 (8.983)	61.100 (9.170)
Median (Q1, Q3)	59.000 (52.000, 67.000)	63.000 (56.000, 70.000)	61.000 (53.000, 69.000)
Min - Max	45.000 - 79.000	45.000 - 79.000	45.000 - 79.000
Missing	0	0	0
BMI			
Mean (SD)	27.555 (4.496)	29.761 (5.215)	28.540 (4.953)
Median (Q1, Q3)	27.000 (24.200, 30.500)	29.400 (26.200, 33.000)	28.200 (25.000, 31.700)
Min - Max	16.900 - 45.400	-10.000 - 48.700	-10.000 - 48.700
Missing	0	0	0
GENDER	1397 (56.5%)	1200 (60.1%)	2597 (58.1%)
B.LINE_SYMP			
Missing	1 (0.0%)	0 (0.0%)	1 (0.0%)
0	2138 (86.5%)	1060 (53.1%)	3198 (71.5%)
1	334 (13.5%)	937 (46.9%)	1271 (28.4%)
Knee Pain (P01KPACT30)	530 (21.5%)	679 (34.1%)	1209 (27.1%)
knee_swell	508 (20.8%)	820 (42.0%)	1328 (30.2%)
diff_upstr	1186 (48.1%)	1290 (64.7%)	2476 (55.5%)
knee_stiff_day_limit			
1	1938 (78.8%)	1317 (66.2%)	3255 (73.2%)

2	263 (10.7%)	273 (13.7%)	536 (12.1%)
3	73 (3.0%)	103 (5.2%)	176 (4.0%)
4	75 (3.1%)	120 (6.0%)	195 (4.4%)
5	109 (4.4%)	177 (8.9%)	286 (6.4%)

A table showing the data completeness is presented in Table 0-2. This is for the complete dataset, of which the analysis in this thesis uses a small subset of features. The times throughout the study where the participants were given the PA view radiographs were at 12, 24, 36, 48, 72 and 96 months after their initial baseline assessment. These subsequent measures of the KL score are then used in the survival analysis work, detailed in Chapter 4.

Table 0-2: Follow Up Visit Summary for the OAI study protocol.

Visit	12m	24m	36m	48m	60m	72m	84m	96m	108m
Clinic Visit	4,293 (90%)	4,082 (85%)	3,925 (82%)	3,831 (80%)		3,239 (68%)		3,117 (65%)	
Telephone Interview	200 (4%)	260 (5%)	349 (7%)	425 (9%)	3,935 (82%)	584 (12%)	3,787 (79%)	531 (11%)	3,204 (67%)
Did not consent to extension					724 (15%)	724 (15%)	724 (15%)	724 (15%)	724 (15%)
Contacted no data		201 (4%)		126 (3%)	39 (1%)	81 (2%)	48 (1%)	83 (2%)	88 (2%)
Withdrew	44 (1%)	97 (2%)	148 (3%)	207 (4%)	NA	50 (1%)	87 (2%)	137 (3%)	332 (9%)
Unable to contact		145 (3%)		137 (3%)	13 (<1%)	6 (<1%)	12 (<1%)	13 (<1%)	31 (<1%)
Deceased	12 (0.2%)	29 (0.6%)	52 (1%)	70 (1%)	85 (2%)	112 (2%)	138 (3%)	172 (4%)	208 (4%)

MOST

The MOST data has 204 variables, containing information on a variety of factors including medication history and questions relating to pain and symptoms, along with demographic information. The information in Table 0-3 shows the retention success in the MOST data collection process during the course of the study. The protocol for the type of data that was collected is also shown in Table 0-3. The clinical data collections comprised of both telephone interviews and clinical visits.

Table 0-3: Follow Up Visit Summary for the MOST study protocol.

Time point	Enrolled[1]	Data collected[2]	Clinical Data	Radiograph Images
Baseline	3026	3026 (100%)	3026	3015 / 3011*
15-Months	3018	3007 (100%)	3007	293
30-Months	2993	2969 (99%)	2969	2651
60-Months	2882	2768 (96%)	2768	2114 / 2100*
72-Months	2778	2715 (98%)	2715	-
80-Months	2721	2638 (97%)	2638	1961

[1] "Enrolled" means enrolled at baseline and continuing participation (not deceased and not withdrawn).

[2] "Data collected" means all or some data collected (measurements and exams completed or partially completed).

Where * denotes the use of full limb radiographs, and the remaining are PA and Lateral View radiographs of the joints.

The variables contained in Table 0-4 are a summary of information relating to the variable subset used for the diagnostic modelling validation in Chapter 5 of this thesis.

Table 0-4: Summary table for the MOST data detailing the variables used within the thesis.

	0 (N=1418)	**1 (N=1571)**	**Total (N=2989)**
BMI			
Mean (SD)	29.184 (4.944)	32.147 (6.474)	30.741 (5.983)
Median (Q1, Q3)	28.720 (25.633, 32.098)	31.090 (27.660, 35.405)	29.875 (26.657, 33.740)

Min - Max	16.720 - 52.390	18.250 - 71.910	16.720 - 71.910
Missing	0	1	1
AGE			
Mean (SD)	60.783 (7.933)	63.945 (7.972)	62.445 (8.107)
Median (Q1, Q3)	60.000 (54.000, 67.000)	64.000 (57.500, 70.000)	62.000 (55.000, 69.000)
Min - Max	50.000 - 79.000	50.000 - 79.000	50.000 - 79.000
Missing	0	0	0
knee_stiff_day_limit			
1	1165 (82.2%)	999 (63.6%)	2164 (72.4%)
2	104 (7.3%)	192 (12.2%)	296 (9.9%)
3	25 (1.8%)	60 (3.8%)	85 (2.8%)
4	36 (2.5%)	72 (4.6%)	108 (3.6%)
5	88 (6.2%)	247 (15.7%)	335 (11.2%)
diff_upstr	1044 (73.7%)	1402 (89.2%)	2446 (81.9%)
Knee Pain (KPACT30)	1077 (85.3%)	1362 (93.2%)	2439 (89.5%)
Gender	827 (58.3%)	969 (61.7%)	1796 (60.1%)
B.LINE_SYMP	615 (71.9%)	1025 (80.3%)	1640 (77.0%)

OActive

The OActive data was used for validating the diagnostic model, described in Chapter 5. The information in Table 0-5 is a summary of the features used in the model validation in this thesis.

Table 0-5: Summary table for the OActive data detailing the variables used within the thesis.

	0 (N=66)	1 (N=131)	Total (N=197)
Data provider			
ANIMUS	0 (0.0%)	0 (0.0%)	0 (0.0%)
HULAFE	66 (100.0%)	2 (1.5%)	68 (34.5%)
UNIC	0 (0.0%)	129 (98.5%)	129 (65.5%)
Gender	22 (33.3%)	34 (26.0%)	56 (28.4%)
Age			
Mean (SD)	51.970 (6.031)	70.061 (8.444)	64.000 (11.517)
Median (Q1, Q3)	52.500 (47.000, 57.000)	71.000 (64.000, 76.500)	64.000 (55.000, 73.000)
Min - Max	41.000 - 65.000	50.000 - 85.000	41.000 - 85.000
Missing	0	0	0

BMI			
Mean (SD)	27.364 (3.627)	29.925 (4.710)	29.067 (4.533)
Median (Q1, Q3)	26.913 (25.101, 29.124)	29.380 (26.950, 32.455)	28.520 (25.970, 31.398)
Min - Max	18.620 - 38.830	20.780 - 46.880	18.620 - 46.880
Missing	0	0	0
Knee_Swell			
No	66 (100.0%)	59 (45.0%)	125 (63.5%)
Unspecified	0 (0.0%)	1 (0.8%)	1 (0.5%)
Yes	0 (0.0%)	71 (54.2%)	71 (36.0%)
kneepain	0	129 (100.0%)	129 (100.0%)