Can the Potential Role of Business intelligence tools (with the application of predictive analytics) identify patients at high risk of not attending Irish Healthcare appointments?

Mustafa Magdeldin

M.Sc In Health Informatics **2018**

A dissertation submitted to Trinity College Dublin

in partial fulfilment of the requirements for the degree of

Master of Science in Health Informatics

Declaration

I declare that the work described in this dissertation is, except where otherwise stated, entirely my own work, and has not been submitted as an exercise for a degree at this or any other university. I further declare that this research has been carried out in full compliance with the ethical research requirements of the School of Computer Science and Statistics.

Signed:_____

Mustafa Magdeldin

17st July 2018

Permission to lend and/or copy

I agree that the School of Computer Science and Statistics, Trinity College may lend or copy this dissertation upon request.

Signed:_____

Mustafa Magdeldin

17st July 2018

Abstract

The purpose of this study was to evaluate if the use of business intelligence tools with the application of predictive analytics, could predict patients at high risk of not attending an appointment in an Irish health care context. The aims of this study were to:

- Use statistical analysis tools, explore relationships between variables associated with missed appointments to help understand the most commonly affecting factors.
- Develop a statistical model with a cut-off threshold that predicts the probability a patient will miss an appointment taking the most common factors into consideration and provide a demonstration of how this was achieved
- Evaluate if it is possible to accurately predict patients at high risk of missing an appointment

Data obtained from a radiology department from an Irish hospital collected over a four-year period (2014-2018) was used for a quantitative approach involving statistical analysis to develop a model that can be used not just for this department but a general model that can be applied to any Imaging department in a hospital or clinic with a scheduling system. Patient appointment details that are available in almost every EHR and scheduling system were used for this study. All patient factors were taken into consideration after literature reviews, analysis and modelling were proven to affect patient non-attendance and targeted intervention steps were proposed to reduce the number of missed appointments.

Acknowledgements

I would like to take this opportunity to thank my supervisors Professor Lucy Hederman and Professor Mary Sharp for their advice and guidance during the writing of my dissertation, without their support this dissertation would not have been possible.

And finally, I would like to thank my family and friends for their continued support and encouragement.

Table of contents

Declaration 2

Permission to lend and/or copy 3

Abstract 4

Acknowledgements 5

Chapter 1 Introduction 15

1.1 Introduction 15

1.2 Business intelligence in healthcare 17

1.3 Research Question and Objectives 17

1.4 Overview of the Research 18

1.5 Overview of the Dissertation 18

Chapter 2 Literature Review 20

2. 1 Introduction 20

2.2 Factors affecting no-shows 20

2.3 Population-Based Models 21

2.4 Individual-Based Models 21

2.5 Statistical Binary Classification techniques 22

2.6 Predictive analytics 23

2.7 Interventions 31

2.8 Conclusion 32

Chapter 3 Methodology 33

3.1 Introduction 33

3.2 Clinical Setting 33

3.3 Methods 34

3.3.1 Data 34

3.3.2 Pre-Processed Data - DNA Dataset 34

3.3.3 Processing of the DNA Dataset 36

3.3.4 Pre-Processed Data - Attended Dataset 39

3.3.5 Processing of the Attended dataset 40

3.3.6 Merging of Both Data Sets 42

3.4 Preliminary Descriptive Statistics of Frequencies 43

3.4.1 Waiting time in days Variable 43

3.4.2 Days of the week Variable 44

3.5 Data Protection 45

3.6 Description of the data 48

3.6.1 Attendance 48

3.6.2 Morning and Afternoon appointments 49

3.6.3 Days of the week 50

3.6.4 Referral Source 51

3.6.5 Modality 52

3.6.6 Males vs Females 53

3.6.7 Patient Class 53

3.6.8 Order Priority 55

3.6.9 Waiting Time in Days 56

3.6.10 Age 57

3.7.1 Attendance variable with respect to numerous variables in the dataset 58

3.7.2 Attendance vs Gender and Modality 58

3.7.3 Attendance vs Patient class and referral source 60

3.7.4 Attendance vs Time of day and day of the week 61

3.7.5 Attendance vs Gender and Order Priority 62

3.8 Conclusion of descriptive statistics 64

Chapter 4 Development of the predictive model 65

4.1 Introduction 65

4.2 Development of the predictive model 65

4.3.1 Output of Binary Logistic Regression Test run 1 66

4.3.2 Variables in the equation test run 1 69

4.4.1 Output of Binary Logistic Regression run 2 72

4.4.2 Variables in the equation run 2 74

4.5.1 Calculating Threshold – ROC Curve 76

4.5.2 Area under the curve 77

4.6 Classification Evaluation with various cut-off points 78

4.7 Conclusion of the development of the predictive model 79

Chapter 5 Interpretation and application of the predictive model 80

5.1 Introduction 80

5.2.1 Results of the model 80

5.2.2 Order priority 82

5.2.3 Modality 82

5.2.4 Referral source 83

5.2.5 Time of day 83

5.2.6 Waiting time in days 83

5.2.7 Age 84

5.2.8 Gender 84

5.2.9 Patient class 84

5.3 ROC curve analysis 84

5.4 Hosmer and Lemeshow test 85

5.5 Practical Application of the Predictive model 85

5.6 Classification Evaluation 88

5.7.1 Target intervention to reduce DNA's 90

5.7.2.1 Communication 91

5.7.2.2 SMS application to missed appointments 91

5.7.3.1 Overbooking 92

5.7.3.2 Overbooking application to missed appointments 93

5.6 Variables under patient control 95

5.7 Limitations of the Study 96

5.8 Conclusion of the results and interpretation 97

Chapter 6 Conclusion 98

6.1 Introduction 98

6.2 Analysis and tools 98

6.3 Key Findings 98

6.4 Recommendations for future research 99

6.5 Contributions to the research 100

6.6 Individual Reflection 100

6.7 Conclusion 101

References 102

List of Tables

Table 2.1. (Daggy et al., 2010) Variables presented 28

Table 2.2. (Daggy et al., 2010) Variables (Continued) 29

Table 3.1 Pre-processed Variable list of Did not Attend dataset 35

Table 3.2 New Variable list of Did not Attend dataset 37

Table 3.3 Snippet of Did not Attend dataset 38

Table 3.4 Pre-processed Variable list of Attended dataset 39

Table 3.5 Snippet of Attended dataset 41

Table 3.6 New Variable list of the Attended dataset 41

Table 3.7 Days of the week Variable 45

Table 3.8 Number of Overall Appointments 45

Table 3.9 Attended vs Did not Attend (Figures) 48

Table 3.10 Morning and Afternoon appointments (Figures) 49

Table 3.11 Days of the week (Figures) 50

Table 3.12 Referral Source (Figures) 51

Table 3.13 Modality (Figures) 52

Table 3.14 Gender (Figures) 53

Table 3.15 Patient class (Figures) 54

Table 3.16 Order priority (Figures) 55

Table 3.17 Central Tendencies (Figures) 58

Table 3.18 Attendance vs Gender vs Modality (Figures) 59

Table 3.19 Attendance vs Patient class vs Referral Source (Figures) 61

Table 3.20 Attendance vs Time of day and day of the week (Figures)..... 62

Table 3.21 Attendance vs Gender vs Order Priority 63

Table 4.1 Case processing summary 67

Table 4.2 Omnibus Tests of Model Coefficients..... 68

Table 4.3 Model Summary 68

Table 4.4 Hosmer and Lemeshow Test 68

Table 4.5 Classification Table 69

Table 4.6 Variables in the Equation 70

Table 4.7 Run 2 Case processing Summary..... 72

Table 4.8 Run 2 Omnibus Tests of Model Coefficients 73

Table 4.9 Run 2 Model Summary 73

Table 4.10 Run 2 Hosmer and Lemeshow Test 73

Table 4.11 Run 2 Classification Accuracy 74

Table 4.12 Run 2 Variables in the Equation 75

Table 4.13 (Hosmer and Lemeshow, 2000)..... 77

Table 4.14 Area Under the Curve 78

Table 4.15 Classification Evaluation..... 79

Table 5.1 Variables in the Equation Results 81

Table 5.2 Maximising the Sum of Sensitivity and Specificity 89

Table 5.3 Classification table for cut-off point 0.3 90

Table 5.4 Schedule Booking Example 94

List of Figures

Fig. 1.1 Patient non-attendance figures according to the HSE 16

Fig. 3.1 Waiting time in Days Variable 44

Fig. 3.2 Data protection Guidelines on research within the health sector 46

Fig. 3.3 Attended vs Did not Attend 48

Fig. 3.4 Days of the week 50

Fig. 3.5 Referral Source 51

Fig. 3.6 Modality 52

Fig. 3.7 Gender 53

Fig. 3.8 Patient Class 54

Fig. 3.9 Order Priority 56

Fig. 3.10 Waiting time in Days 57

Fig. 3.11 Age 57

Fig. 3.12 Attendance vs Gender vs Modality 59

Fig. 3.13 Attendance vs Patient class vs Referral Source 60

Fig. 3.14 Attendance vs Time of day and day of the week 62

Fig. 3.15 Attendance vs Gender vs Order Priority (Figures) 64

Fig. 4.1 Flow of Modelling Process 66

Fig. 4.2 ROC curve 76

Abbreviations

EHR	Electronic Health Record
DNA	Did not attend
BI	Business Intelligence
GP	General Practitioner
MRI	Magnetic resonance Imaging
CT	Computed Tomography
PET	Positron Emission Tomography
ROC	Receiver operating characteristic
PPV	Positive Predicted Value
NPV	Negative Predicted Value
Df	Degrees of freedom
B	B coefficient
Exp(B)	Odds Ratio
Sig	Statistical Significan

Chapter 1 Introduction

1.1 Introduction

Hospitals and clinics operate in a regulated industry of which quality of its performance is evaluated in terms of services provided to patients and the effectiveness of the healthcare provider's process. To be able to sustainably deliver quality care, hospitals need to monitor the care they deliver by measuring various outcomes of patients and to be able to continuously improve processes. Improving the quality of care requires a comprehensive, multifaceted approach to identifying and disseminating the learning from good quality care and from identifying and managing poor quality care in individual services and finding long term solutions. Described in the Irish Health Stat online performance information tool, two of the main performance indicators, are: 1) access, which is described as waiting times in outpatient clinics and planned procedures along with 2) resources, which relates to staffing and budgeting. All these factors have significant impacts in revenue, cost of care and underutilised medical resources and reduced clinical efficiency (George and Rubin, 2003).

A major obstacle in cost effective healthcare delivery and patient safety is non-attendance by patients at appointments. A DNA or no-show is defined as, when a patient does not arrive at a previously scheduled appointment without any prior notice. These anticipated and unpredictable occurrences often lead to delays and unproductiveness which is almost unrecoverable for hospital staff and doctors. Non-attendance constitutes a challenge for all clinicians and hospital administrative staff in all hospitals. (Mbada *et al.*, 2013) has shown the effects down the line for missed appointments reduces the quality of care to patients and increases patient suffering, believed to adversely affect their treatment outcome, recovery time and quality of life. Disruptions to hospital processes and workflow and negative clinician attitudes also hinder the delivery of the quality of care and cause premature discontinuation of clinical services alongside wasting of human and administrative resources.

In the Irish healthcare setting, the Health Service Executive in 2015 determined for 1 out of 6 hospital appointments made in the Irish health system, patients were recorded as DNA. These figures are approximately 488,000 appointments that were recorded as DNA. First-time appointments were recorded at 923,000 of which 139,000 (15%) failed to show up for appointments. Out of 2,375,000 return appointments, 349,000 failed to show up for a follow-up appointment (Fig 1.1). Overall this gives a 15% no-show rate for patients in Irish hospitals.

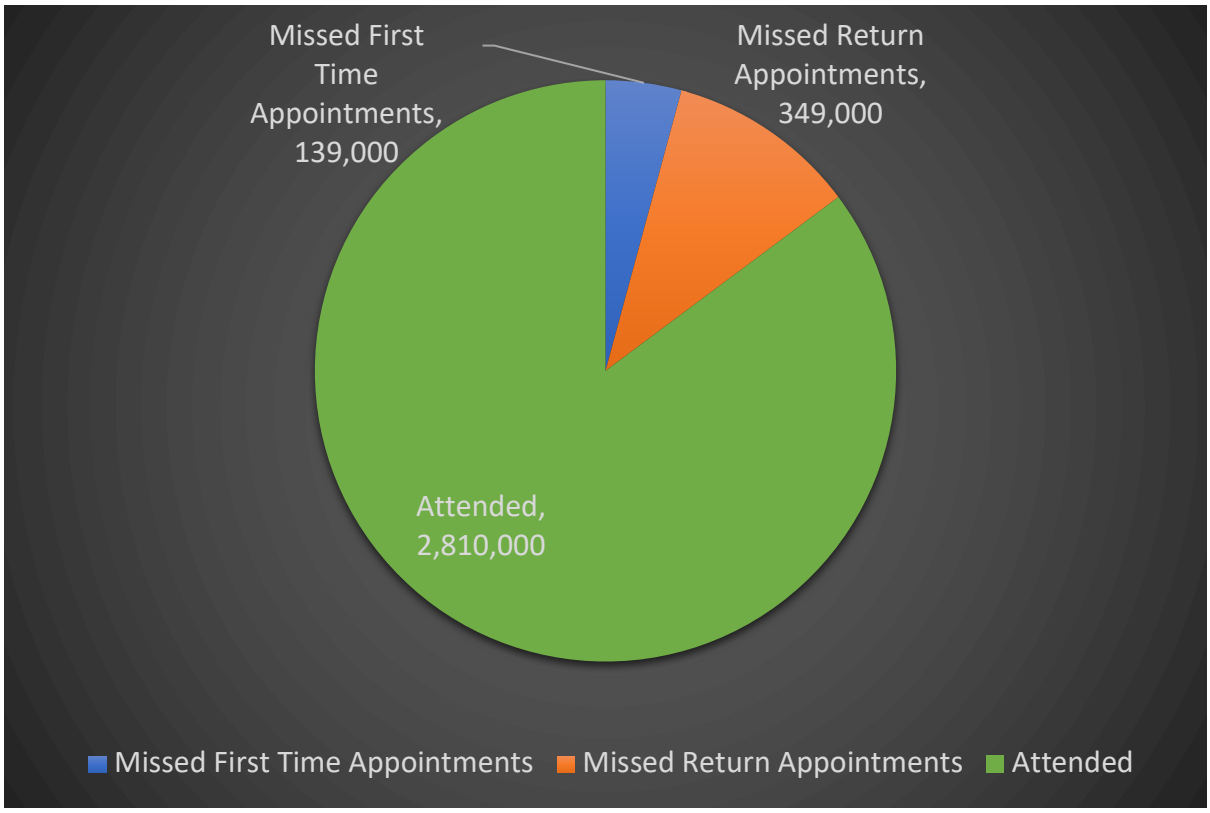


Fig. 1.1 Patient non-attendance figures according to the HSE

The HSE estimates that the average cost of scheduling an appointment is €129 per patient. The cost of a no-show was put at €44 arising from reviewing referral notes as well as communication back and forth between GP's and obtaining medical records. Putting all this into perspective the cost of "did not attend" appointments equates to €21.5 million a year using the average cost of a no-show appointment and the recorded number of patients who did not attend. However, these are general statistics which include appointments of outpatient services throughout Ireland.

1.2 Business intelligence in healthcare

Hospital management cannot effectively work without the utilising of key performance indicators to control business procedures, therefore BI is a must in the health sector. Budget cuts in healthcare and higher demand from patients requires the ability to react quickly, change strategies and optimize processes. All the above actions need information and data readily available to tackle these issues. Therefore, the concept of business intelligence applications comes into play and opening opportunities to challenge these tasks. Predictive capability is the ability to build and analyse a model for making accurate predictions of observations that can be taken into consideration into planning and making estimations. Prediction of future trends through the extraction of large datasets and identifying trends is all is all part of predictive analytics and applying algorithms such as regression analysis.

1.3 Research Question and Objectives

While there is current momentum and support for the establishment of business intelligence capabilities in the healthcare sector and despite the increasing data held electronically there is a lack of established frameworks for using business intelligence tools to measure various outcomes and trends. In this study, the use of non-uniform and inconsistently structured data will be used for the objectives defined. The research question for this study is: Can patients at high risk of not attending a radiology appointment be predicted from the data exported and combined from an EHR and scheduling system?

The objectives of this study are:

Evaluate the characteristics and trends of patients that have missed appointments specifically relating to a large Irish Hospital radiology department with the aid of business intelligence tools using historical scheduled appointments.

1. Use statistical analysis tools, explore relationships between variables associated with missed appointments to help understand the most commonly affecting factors.

2. Develop a statistical model with a cut-off threshold that predicts the probability a patient will miss an appointment taking the most common factors into consideration and provide a demonstration of how this was achieved
3. Evaluate if it is possible to accurately predict patients at high risk of missing an appointment

1.4 Overview of the Research

The research question was addressed through a quantitative statistical analysis of the dataset obtained from the imaging department of an Irish hospital. The overall dataset contained 24,879 appointments from November 2014 to October 2019. Data was pre-processed to fit binary logistic regression analysis, which was used to determine patients that are high risk of missing a scheduled appointment. This analysis measured the probability of a patient missing an appointment taking into consideration the characteristics of the patient and previous missed appointment history, giving an outcome of true or false.

First a literature review was conducted to evaluate methods used in previous research. The subject of statistical analysis and predictive modelling was the primary focus. Most studies used a common variable set that was available in most EHR and scheduling systems for hospitals and clinics

1.5 Overview of the Dissertation

This chapter presented the motivations for the research, research question and objectives, a background to the topic of missed appointments.

Chapter 2 provides a literature review of the methods used to predict patient non-attendance. Firstly, the impact of business intelligence tools in healthcare with the focus of

predictive analytics is defined followed by a detailed insight into the research in previous studies.

Chapter 3 presents how the data was pre-processed to suit binary logistic regression. A detailed description of what the data consists of will be presented and variables combined to understand the data which relates to attended and DNA appointments.

Chapter 4 presents the development of the predictive model for the classification of attended and did not attend appointments.

Chapter 5 presents the model, and assesses its accuracy and usefulness. It then discusses how the results address the research question, the significance of the results and the limitations of the study. A section for target interventions that can reduce Patient non-attendance will be presented.

Chapter 6 Presents a conclusion with a review of the research and its findings and an individual reflection on the topic. Future work in this area is also outlined.

Chapter 2 Literature Review

2. 1 Introduction

A review of the literature conducted seeks to put the topic researched into the context in terms of other work that has been carried out in this area. The purpose is to describe the relationship of each work to the other under consideration and identify new ways of understanding and interpreting other methods of research. It aims to resolve gaps in previous research conducted and settle conflicts amongst other studies in the area. It will also help to identify the need for research in other areas. This study has focused on search terms such as data analytics, business intelligence tools, patient no-show, missed appointments and patient missed appointment interventions. Several databases were used for this research; Scopus, Medline, Google scholar and Science direct. The following journals were also used; International journal of Medical informatics, Health Informatics journal and the Journal of the American medical informatics journal. Relevant articles were also searched through google.

2.2 Factors affecting no-shows

Several studies have discussed the individual factors that contribute to missed appointments. Personal patient information such as age, sex, gender and nationality have been studied. Many of these variables are readily available in scheduling systems and EHR systems. (Bean and Andrew, 1995) looked at the percentage of appointments missed by age and gender, contrary to expectations based on previous and most recent research, differences among age groups were not significant. However different studies have conducted research in various hospitals with different clinical populations, for example primary care or chronic care, which can have a major effect on characteristics of missed appointments. (Huang and Hanauer, 2014) conducted a study with a data set from a paediatrics clinic and took into consideration prior no-show history and found that it played a significant factor. They also concluded that gender and religion did not play a role in patient no-shows. (Daggy *et al.*, 2010) investigated data from a U.S. Veterans clinic and took into consideration the patients' clinical condition such as; cardiac condition, major depression, stroke or dementia, chronic pain and congestive heart failure. These factors affected the patient no-show rate. 77% of the patients in the

dataset were greater than 50 years of age and patients with a cardiac condition are least likely to miss an appointment, compared to patients with depression and drug dependency issues. A few studies have also considered the effects of personal issues such as oversleeping and forgetting. (Neal *et al.*, 2005) considered personal reasons why appointments are missed in general practices in the UK. Forgetfulness was the main reason for missed appointments along with family commitments, being too ill to attend and common misunderstandings or mistakes

2.3 Population-Based Models

Statistics and machine learning play an important role in population-based techniques for analysis which will be demonstrated for predicting missed appointments in this dissertation. These methods use information from a whole population dataset in the form of factors and variables to estimate the probability of a patient showing up to a scheduled appointment. Logistic regression is one of the most popular statistical methods in this category that is used for binomial regression, which can predict the probability of a no show by fitting numerical or categorical predictor variables (Dove and Schneider, 1981). (Glowacka, Henry and May, 2009) conducted research using tree- and rule-based models, which create if–then constructs to separate the data into increasingly homogeneous subsets, based on which the desired predictions of no-show can be found. The problem with using a tree is even a small change in the input parameters can at times cause large changes in the tree leading to instability. Any irrational expectations in the decision tree can lead to flaws and errors and only follows a natural course of defined outputs. Once the decision tree model has been built adding data can have a very small effect.

2.4 Individual-Based Models

Individual based approaches are based on the global consequences of local interactions of members of a population. They are primarily time series and smoothing methods that are used for no-show prediction. The characteristics of individuals are tracked through time. (Alaeddini *et al.*, 2015) Time series methods forecast future events based on past events by using stochastic models. Individual based methods usually employ a function of past data

(attendance record) to estimate the probability of a future event, e.g. no-show and cancellation. But for the initial state where there is no previous history available, such a function is inapplicable, and consequently, individual-based methods usually use random guess for the initial estimate, e.g., probability of no-show.

2.5 Statistical Binary Classification techniques

Machine learning has been used for prediction in many areas to predict behaviour and outcomes within the medical field. In the area of business, prediction is also used to aid business executives making decisions, for example identifying consumer preferences for products. There are various tools and software to help in these predictions that will be used in this dissertation. A number of different methods exist, which include support vector machines and neural networks. Decision trees, random forest and logistic regression all provide classifications of outcomes. Logistic regression will be the focus in this study.

Random forest is a learning method for classification, that consist of a multitude of decision trees (Breiman, 2001). This approach has several advantages over other methods of statistical modelling. There is no need for a predefined hypothesis making it less likely to overlook potential interactions in the model.

Support vector machine is another discriminative classifier which sorts the data into one of two categories. In this algorithm, each data item is plotted as a point in n-dimensional space (where n is the number of features you have) and then classification is performed by finding the hyper-plane that differentiates the two classes. They are simply the coordinates of each individual observation separated (Wang, 2005).

Among many epidemiological studies logistic regression has been the go to method for predicting outcomes in the field of statistics. The regression calculates values for the coefficients for each input variable. Logistic regression results in a formula (or model). When the formula is applied to a case to be classified, the value obtained is the predicted probability that the case belongs to the target class (e.g. DNA). Applications can then decide what threshold predicted value to use, how sensitive or specific they want the test to be. For example, if the predicted value is greater than 0.5 classify this entry as a DNA. This adds flexibility to the model and thresholds can be set. Another reason why logistic regression was

chosen over other methods, there have been many studies around the world conducted around missed patient appointments and there is already a hypothesis in existence of various predictor variables that contribute to patient missed appointments.

2.6 Predictive analytics

Much of the research conducted around patient DNA rates has revolved around the use of predictive modelling and analytics to try and predict patients that are most likely to miss an appointment while considering many variables that are accessible in most electronic health records and scheduling systems. This gives way to the area of healthcare analytics alongside business intelligence, where big data is available. An important question to consider when using the concept of healthcare analytics is how can this affect the end user's knowledge? (Ivan and Velicanu, 2015) stated that the power of collective insights is achieved through the process of analytics and is realised in the following three steps:

- Engagement: Predict demand and supply of the supply chain
- Visualise: Understand the customer's thoughts
- Predict: Provide the proper offers and services to every customer; also predict new markets and trends and innovate new products

The increasing demand in modern healthcare today, is to develop a deep understanding of efficiency and quality of care they deliver. Using the power of collective insights through the steps mentioned, being able to predict patients that are at high risk of not attending an appointment for financial reasons or quality of care, understanding the characteristics of these patients that tend to miss appointments and with these steps used for the future of healthcare new products and services can come about. These new products and services can be incorporated into advanced scheduling systems and integrated into EHR systems or can lead to better processes and workflows.

Controlling business procedures and workflows is a result of analytics, and along with this brings the use of business intelligence tools to help define key performance indicators. (Luhn, 1958) defined business intelligence *"as a collection of activities carried on for whatever purpose, be it science, technology, commerce, industry, law, government or defence. The communication facility serving the conduct of a business (in the broad sense) may be referred to as an intelligence system".* The notion of intelligence is also defined here, in a more general sense, as *"the ability to apprehend the interrelationships of presented facts in such a way as to guide action towards a desired goal".* Theoretical studies of business intelligence are categorized into two different types. (Negash, 2004) highlighted these as: 1) the concept of a data centre and advocated the use of BI to combine *"operational data with analytical tools to represent complex and competitive information to planners and decision makers to improve data quality and timeliness of decisions during decision making".* 2) The emphasis of an idea of a process centre, in which organisations themselves are viewed as "*process integration*". Only the first concept of operational data with analytical tools will be applied as it directly relates to this dissertation.

(Devasahay, Karpagam and Ma, 2017) analysed 12 months of data for 2013 for a number of clinics in Singapore. The overall no show rate was at 18.59%. Variables such as distance from hospital replaced addresses to anonymize the data. As in many DNA studies the first observation that was made and calculated was gender and age. The study was started with an assumption that age and gender played a major role in missed appointments (according to hospital predictions). Preliminary statistics calculated showed that males had a higher tendency to miss appointments which was later proved in the analysis. Age was analysed but there was no correlation with higher missed appointments due to age. Other variables that were taken into consideration were "visit time" and "visit date". These time related variables were broken down by month and day, along with 3 other time of day values relating to morning, afternoon and evening. Morning and afternoon appointments were found to have more regular appointment attendances and more misses tend to happen in the evening. Errors in the data set led researchers to conclude there was incorrect capture by the system on Sundays. Public vs private patients were also studied and there was a higher rate of missed appointments with public patients. Logistic regression and decision trees were the models

used to predict appointment misses. It was found that decision trees were the more accurate method of calculating. The sensitivity, which is the probability the test will predict a missed appointment was at 23.22% with a predicted positive value of 15.58% for decision trees. Compared with logistic regression results with a sensitivity of 0.11% and a positive predicted value of 61.68%.

(Huang and Hanauer, 2014) studied 10 years of data consisting of 7,988 patients and 104,799 visits obtained from a scheduling system and electronic health records in a paediatrics clinic from January 2002 – December 2011. The study looks at predicting patient no-show with the additional layer determining how effective over booking appointments can be, which is discussed in more detail in the Interventions section. The dataset was selected upon availability in the EHR system. The first set of variables included distance from hospital, visit type, time to appointment, day, month and count of prior missed appointments. The second category was under demographic information which included age, language, religion, race and gender. The third category was based on their health insurance, if they were covered or not. They defined the definition of no-show as not arriving to the appointment and not giving warning to the hospital. In other literature reviews a no-show could have included patients giving a warning before cancelling. Statistics showed that 11.2% were no-shows overall in the 10 years of the data.

The threshold of a no-show has been studied in previous literature and covered by (Huang and Zuniga, 2012) to support the output of true or false. This threshold favours doctors and administrative staff time and takes costs of idle time into consideration. Applying this threshold calculated to (Huang and Hanauer, 2014) the validation dataset, among 19,871 visits, the model successfully predicts 17,104 cases, which translates to 86.1% of scheduled visits and a predicted no-show rate of 13.9%. Using logistic regression, a hypothetical scenario was constructed to show the probability of a missed appointment:

A fifteen-year-old, English-speaking African American adolescent, with 6 people in her household and a history of five prior no-shows to the clinic. She made the appointment about

two months in advance for a routine return visit, scheduled for a Friday afternoon in March. She lives in a county that is adjacent to the county in which the clinic resides, about 22 miles away. She has more than one insurance carrier, but her primary insurance is HMO under herself.

The set of variables included in this model were:

- Visit type
- Age
- County to clinic location
- Distance to the clinic
- Appointment time
- Appointment day
- Appointment month
- Time to appointment
- Race
- Gender
- Religion
- Language
- Total insurance carriers
- Primary insurance
- Main insurance holder
- Number in household
- Count of prior no shows

The model predicted that for the above scenario the patient has a 79% chance of missing her appointment taking the above considerations. According to the Chi-square test results it was concluded that neither gender nor religion was a contributing factor to missing appointments and therefore these were removed.

(Daggy *et al.*, 2010) investigated the modelling of patient no-show rates using logistic regression and used the estimates to improve scheduling at a hospital to maximise clinic times while maintaining short patient wait times. The data collected over a three-year period consisted of 32,394 visits from 5,446 patients. No show results for this data set was considered accurate as the data was entered manually by a clerk on the day of the no-show. Much of the literature failed to state how the data was entered to ensure its validity. The data set was randomly split into 2 groups. One was for the testing and the second was the validation data, respectively two thirds and one third. Patients with only one appointment were removed as the major predicting variable was previous patient history. (Huang and Hanauer, 2014) also considered this as a large contributing factor to accurate prediction. The data set in this literature was broken down in more detail. Age variables were split into 4 categories, ≤ 50, 51-50, 61-70 and <70. Other variables included were: marital status, insurance type and distance to hospital. Distance to hospital was categorised into ≤ 6 miles, 7-90 miles and <90 miles. Appointments were broken down by weekday and AM or PM. A true or false variable was assigned to appointments that were scheduled within 14 days. Clinical characteristics were also included such as; cardiac condition, major depression, stroke or dementia, chronic pain and congestive heart failure. The variables are represented in Tables 2.1 and 2.2.

Table 2.1. (Daggy et al., 2010) Variables presented

N = 3484	Last appointment no-show	
	Yes N (%)	No N (%)
Demographic characteristics		
Age (years)		
≤ 50	249 (31.4%)	543 (68.6%)
51–60	169 (17.2%)	816 (82.8%)
61–70	62 (9.4%)	598 (90.6%)
> 70	56 (5.3%)	991 (94.7%)
Marital status		
Married	183 (9.5%)	1739 (90.5%)
Single/widowed/divorced	352 (22.7%)	1202 (77.3%)
% costs covered by VA		
< 20%	422 (16.4%)	2157 (83.6%)
20–60%	84 (13.2%)	553 (86.8%)
> 60%	29 (11.2%)	231 (88.8%)
Insurance		
Medicare or private	347 (16.0%)	1828 (84.0%)
Other insurance	188 (14.5%)	1113 (85.5%)
Distance to VA		
≤ 6 miles	183 (21.1%)	684 (78.9%)
7–90 miles	297 (12.0%)	2172 (88.0%)
> 90 miles	48 (35.3%)	88 (64.7%)
Appointment characteristics		
Day of week		
Monday	98 (14.4%)	581 (85.6%)
Tuesday	123 (14.4%)	734 (85.6%)
Wednesday	62 (14.4%)	369 (85.6%)
Thursday	117 (15.6%)	634 (84.4%)
Friday	136 (17.8%)	630 (82.2%)
AM appointment		
Yes	270 (15.2%)	1507 (84.4%)
No	266 (15.6%)	1441 (84.8%)
Scheduled within 14 days		
Yes	64 (12.1%)	463 (87.9%)
No	472 (16.0%)	2485 (84.0%)
Winter		
Yes	154 (24.9%)	464 (75.1%)
No	382 (13.3%)	2484 (86.7%)
Clinical characteristics		
Charlson Index		
0	372 (18.1%)	1682 (81.9%)
1	101 (12.1%)	732 (87.9%)
≥ 2	63 (10.6%)	534 (89.4%)
Hospital admissions		
0	400 (14.5%)	2353 (85.5%)
1	73 (18.0%)	333 (82.0%)
≥ 2	63 (19.4%)	262 (80.6%)

Table 2.2. (Daggy et al., 2010) Variables (Continued)

N = 3484	Last appointment no-show	
	Yes N (%)	No N (%)
No. previous scheduled visits		
≤ 3	232 (20.6%)	892 (79.4%)
4–6	172 (13.3%)	1118 (86.7%)
> 6	132 (12.3%)	938 (87.7%)
Diabetes		
Yes	106 (10.8%)	873 (89.2%)
No	430 (17.2%)	2075 (82.8%)
Cardiac condition		
Yes	72 (7.3%)	909 (92.7%)
No	464 (18.5%)	2039 (81.5%)
Major depression		
Yes	144 (19.6%)	592 (80.4%)
No	392 (14.3%)	2356 (85.7%)
Stroke or dementia		
Yes	24 (16.1%)	125 (83.9%)
No	512 (15.4%)	2823 (84.6%)
Chronic Pain		
Yes	313 (15.2%)	1744 (84.8%)
No	223 (15.6%)	1204 (84.4%)
Congestive heart failure		
Yes	31 (11.0%)	250 (89.0%)
No	505 (15.8%)	2698 (84.2%)
Chronic obstructive pulmonary disease		
Yes	109 (13.2%)	719 (86.8%)
No	427 (16.1%)	2229 (83.9%)
Drug dependence		
Yes	237 (23.6%)	769 (76.4%)
No	299 (12.1%)	2179 (87.9%)
Use narcotics		
Yes	201 (16.8%)	994 (83.2%)
No	334 (14.6%)	1947 (85.4%)
Continuous characteristics	*Mean (SD)* *N* = 536	*Mean (SD)* *N* = 2948
Days since last scheduled visit	154.7 (114.8)	169.6 (106.5)
No. hospital admissions	0.6 (1.7)	0.4 (1.1)
Prior no-show rate	0.3 (0.3)	0.1 (0.2)

The process for validating the results was performed on the estimated no-show probabilities calculated. The validation method used here was the Monte Carlo simulation (Law, 1991). First, 1000 samples of size 30 (the average number of patients seen daily) from 3,631 patients were randomly selected (with replacement) from the validation cohort of 1,815 patients. Then for the given sample the number of unexpected no-shows was computed and compared to the actual number of no-shows. The tests revealed that age was a factor, the young and non-married patients with less medical costs covered were most likely to miss appointments. Patients living within 6 miles had a no-show of 21.1% while those living 7-90 miles had a 12% no-show rate and greater than 90 miles had a 35.3% rate. It's interesting to see that the 7-90-mile range had the least rate of no-show, this can be where more seriously ill patients are living and need to visit the hospital more often for other services. Appointments with a lead time of more than 2 weeks were more likely a no-show as were appointments in winter. Patients with more serious health issues such as cardiac conditions, diabetes and congestive heart failure were less likely to miss an appointment, whereas those with drug dependencies and depression were more likely to miss appointments. However, a limitation of this study is that all patients were from a Midwestern Veteran Medical centre and all patients were male. In previous literature, being male was a vital contributing factor to missed appointments.

As most of the literature that was studied was based on quantitative methods, search results did not return much qualitative studies. Another quantitative investigation in the UK by (Neal *et al.*, 2005) considered the reasons why missed appointments in general practices in the UK occurred. This was in the form of a postal questionnaire and was focused on personal patient behaviour. The study was conducted over a three-week period in 2001, in seven practices in West Yorkshire. The seven practices all differed in size, patient count and workload. Of the 386 patients who missed an appointment only 122 (32%) responded.

Over 40% of those who responded to the questionnaire said they forgot about the appointment. Personal behaviour that was also a factor were family commitments, being too ill to attend, misunderstandings or mistakes.

2.7 Interventions

Literature with regards to interventions for reducing patient missed appointments were also studied. Different communication methods were assessed and compared with each other. The main areas of intervention for patient missed appointments was effective communication, whether it be by SMS texting or telephone calls, in various types of hospitals and clinics. In previous studies by (Moberly, 2014) patient appointment reminders with the use of SMS text messaging were investigated within 114 NHS providers in London and resulted in a reduction in 8.8% of DNA's from 2008 to 2009 and 2012 to 2013. In another study by (Brannan *et al.*, 2011) in Scotland at an ophthalmology clinic, patients were sent appointment reminders vis SMS with the aim of specifically reducing DNA's. The non-attendance rate compared with historic non-attendance rate was recorded. Two hundred and one patients were recruited. The historic DNA rate was 12%. The DNA rate in the SMS text reminder group was reduced to 5.5%. They concluded that routine SMS texting is a cost-effective way of reducing DNA's and should become a standard practice. In addition to this a 2-way messaging system could allow for further efficiency for cancellations and rescheduling. The effectiveness of telephone calls vs SMS text reminders was assessed in a separate study. As telephone calls are the more costly and resource full method of communication a randomised controlled trial was investigated by (Junod Perron *et al.*, 2013) in the primary care division of a Geneva university hospital between November 2010 and April 2011. A cost-effective comparison was studied and they found that telephone reminders were at a cost of €0.08 per phone call compared to SMS text messaging which was at a cost of €0.07. They concluded that SMS text messaging reminders were just as effective as a telephone call for decreasing the rate of missed appointments in patients. (Atherton *et al.*, 2012) discussed the advantages and disadvantages of coordinating appointments via email. Where email systems are in place they are most suitable for non-urgent situations as there can be delays in replies between patients and hospital administration. They are well suited to reminders and potentially can be automated at no extra cost requiring minimal human intervention for the simplest reminders. Potential downsides to email were also discussed and privacy concerns were raised along with the digital divide of the elderly, non-English speakers and lower income groups.

Overbooking was also a popular method for reducing appointments in previous literature. Specifically, how predictive modelling can help with effective overbooking approaches. (Huang and Zuniga, 2012) looked at dynamic overbooking scheduling systems to improve patient access while focusing on the individual patient no-show probability. This study did not look at patients' individual needs but rather focuses on physician's idle time, cost and patient waiting time. The proposed approach in the mentioned paper was that each clinic determines the least number of patients to schedule without overbooking and then uses the over-booking approach to find where and how many to overbook based on the objectives of minimizing the total cost including the costs of patient wait time, physician idle time and overtime. They calculated the no-show probability threshold that minimizes the total cost to determine the total overbooked patients.

Literature that showed up in searches have suggested advanced booking schedules allowing the patient to choose their own times though web based appointment systems. (Siddiqui and Rashid, 2013) reported that non-attendance rates for appointments booked by patients online at a dermatology clinic were much lower (6.9%) that the no-show rates of appointments made by traditional means of contacting the clinic and waiting for appointments to be scheduled (17-31%). (Walters and Danis, 2003) also reported that the use of an online web-based communication tool reduced no-shows by 42%. This reduction in missed appointments was implemented for 650 providers in Northern New England in the United States and consisted of a clinical messaging service that let patients request, review, reschedule and cancel appointments. In the UK a study by (Parmar *et al.*, 2009) attendance rates at an audiological medicine clinic, using a "choose and book system" reported higher attendance rates than traditional appointment booking methods.

2.8 Conclusion

The main purpose of a literature review is to demonstrate the topic of predictive analytics and how it can relate to patient non-attendance and to show where it fits into the terms of the wealth of knowledge already gathered on this topic. The aim of this study is to establish whether it is possible to accurately predict patients at high risk of not attending appointments with the application of predictive analytics by applying previous methods and techniques .

Chapter 3 Methodology

3.1 Introduction

This chapter will discuss the processing of the data in detail that will be needed for the analysis. A section describing the variables will be included here and a detailed write up will be included of how the data was processed including a section on data privacy and protection. This chapter will also include the process of data clean-up and standardisation of the dataset to suit the statistical analysis methods used.

3.2 Clinical Setting

The study site where the data for this research was obtained is an Irish hospital. It is considered as a referral centre and the radiology department deals with complex cases from every county in Ireland and provides a complete digital diagnostic imaging service to patients. A service is also provided to GP's in the catchment area. Services include CT scanning, General X-rays, Mammography and breast imaging, MRI Scanning, Ultrasound, PET, CT scanning and nuclear medicine scanning. The department consists of public and private clinics depending on the patient's healthcare cover.

According to 2011 consensus data, where the hospital is located consists of a population of about 1,273,000 encompassing an area of about 115 square kilometres. Along with serving Dublin as the main county, Ireland has a population of around 4,600,000 million, all of whom also have access to the imaging facilities listed above.

3.3 Methods

3.3.1 Data

Data was obtained from a hospital based in Ireland. Approval for the data obtained was granted by the hospital. No ethical approval was needed from Trinity College Dublin as there was no involvement of patients or other people. The data is an extract of the electronic ordering system used by the hospital. The order is placed by the doctor, typically working for the hospital. The order is then processed and scheduled by the clerical officer from the department. The data consisted of 24,879 appointments collected over a four-year period from 2014 to 2018. The data came in two sets, with different fields. The first dataset consists of appointments not attended (DNA dataset) over a period of 3 years; the second set consists of appointments attended (attended dataset) in early 2018. The data concerns appointments for various scans consisting of Magnetic Resonance Imaging, Computed Tomography, Mammograms and Ultrasounds. For each appointment record, there is an indication as to whether the patient attended or missed the appointment.

Prediction of DNAs involves distinguishing DNAs from attended appointments. This requires that the data about DNAs be comparable to data about attended appointments. Thus, the two, separate and different datasets needed to be aligned, with the same columns of data in the same form.

The two datasets and their pre-processing are now described in detail

3.3.2 Pre-Processed Data - DNA Dataset

The initial raw data of DNA appointments obtained ranges between November 2014 to December 2017 and consisted of 18,569 appointments in Excel format. There was a total of 25 fields in the DNA dataset, of which included the medical record number the hospital assigned to each patient. This was converted to a pseudorandom number. Names of patients had been removed and date of birth had been converted into an age before the data set was obtained. Other fields included scheduled date, appointment request date, gender, age, address, imaging modality, patient class (public or private), referral source and other internal

hospital codes for identifying specific procedures and scans. Table 3.1 lists the variables included in the dataset before any changes were made.

Table 3.1 Pre-processed Variable list of Did not Attend dataset

Variable name	**Description**
Original order date and time	Date appointment requested and time
Order date	Date appointment requested
Scheduled date	Appointment date
Scheduled date and time	Appointment date and time
Request Date	Appointment date
Requested for date	Date and time appointment requested
Referral Source	Source of referral
Order physician Alias	Physician that requested the referral
Modality	Type of imaging scan (MRI, CT, US, MMG)
Pseudo MRN	Patient Unique Identifier (coded)
Address 1	Address line 1
Address 2	Address line 2
Address 3	Address line 3
Address 4	Address line 4
Postal Code	Area post code

Age	Age
Gender	Gender
Order Location	Hospital the referral came from
Patient class	Public or private
Exam description	Area to be scanned
Order priority	Prioritised appointment status (Urgent, planned follow-up, routine, clinical trial)
Scheduled status	Appointment stage
Scheduled appointment date	Appointment Date
Order to days	Date appointment was ordered to appointment date
Order comment	Patient comment section (reason for DNA)

3.3.3 Processing of the DNA Dataset

To standardise the data into a structured format fields had to be removed and some added. Items which have no bearing on a patient's likelihood of attending and appointment are irrelevant and are removed. The physician requesting the order was an alias name was also removed. The postal code was also removed as it was an internal code for the catchment area. The exam description column was removed as there was no standard to how that column was entered and it consisted of several exams per patient. The order comment column consisted of comments marking the patient as did not attend and reasons as to why. In this column, I assigned 1 to appointments that were missed which was every appointment in this dataset.

The data was further anonymised by converting the patient address, which consisted of four columns, to distance from the hospital in kilometres. Google Maps API was used for this. Where Google could not calculate the distance due to spelling errors and wrong entry of data, an error code was entered in this distance field.

New variables such as waiting time in days were calculated which was the difference between the appointment request date and the scheduled date. New variables created are presented in table 3.2, highlighted.

Since the dataset included historical appointments and had a lot of errors and incomplete data, this reduced the dataset from 9,451 to 2,212 appointments. Appointments with invalid distances, incomplete fields and spelling errors were removed. This was essential to increase accuracy and remove incomplete entries for better accuracy when conducting the analysis. A snippet of a scheduled appointment with populated variables is shown in table 3.3.

Table 3.2 New Variable list of Did not Attend dataset

Variable name	**Description**
Pseudorandom MRN	Patient Unique Identifier
Appointment requested on	Date appointment requested
Appointment date	Appointment date
Appointment time	Appointment time
Appointment day	**Appointment day**

Appointment month	**Appointment month**
Morning	**Morning appointment**
Afternoon	**Afternoon appointment**
Waiting time	**Wait time from date appointment requested to scheduled date**
Referral source	Source of referral
Modality	Type of imaging scan
Distance	**Distance from Home to hospital**
Age	Age
Gender	Gender
Patient class	Public or private
Order priority	Prioritised appointment status
Scheduled status	Appointment stage
Did not attend	Appointment attended or missed

Table 3.3 Snippet of Did not Attend dataset

pseudo mrn	appointment requested on	Appointment date	Time	Appointment day	Appointment month	Time of day	Waiting time in days
1477518	30/09/2017	13/02/2018	09:00	Tue	Feb	1	136

Modality	Distance	Age	Gender	Patient Class	Order Priority	Scheduled Status	Scheduled Appt Date	Did Not Attend

Computed Tomography	12.6	59	F	PUBLIC	Routine	Scheduled	13/02/2018	1

3.3.4 Pre-Processed Data - Attended Dataset

The initial dataset of attended patients that was obtained consisted of 6,310 imaging department appointments that were attended during January and February 2018. It included 13 different fields that were exported from the hospital scheduling system. This also included a patient identifier number which was randomly generated by the hospital. Therefore, there was no concept of returning patients in the dataset. It included patient age, order request date, appointment date, gender, imaging modality, patient type (public or private) and referral source. Table 3.4 lists the variables before any changes were made.

Table 3.4 Pre-processed Variable list of Attended dataset

Variable Name	**Description**
Pseudorandom MRN	Patient Unique Identifier
Exam Complete Date/Time	Date Examination was completed on
Patient Location at Exam	Location of Examination
Order Procedure	Area to be scanned and modality
Section	Type of scan
Ordered Date/Time	Date and time appointment requested
Appointment Date/Time	Appointment Date/Time
Schedule Confirmation Date	Date of confirmation of appointment
Schedule Confirmation Day	Day of confirmation of appointment

Schedule Confirmation Month	Month of confirmation of appointment
Age	Age
Gender	Gender
Order Priority	Prioritised appointment status
Patient Type	Public or private/Referral Source

3.3.5 Processing of the Attended dataset

To standardise the attended dataset to match the DNA dataset variables were created accordingly. Items which have no bearing on risk of attendance, such as the exam complete date/time variable and patient location at exam were removed. The order procedure was also removed as this field had various comments with no standards and would not have suited the binary logistic regression model. The section variable was renamed to modality and the abbreviations of the scans were assigned accordingly as Ultrasound, Magnetic Resonance Imaging, Mammography and Computed Tomography. The order date/time was changed to the appointment requested date. Schedule Confirmation date, day and month were removed as these variables were not going to be included in the analysis and didn't match up with the DNA dataset. Age, gender and order priority were kept as is.

New variables were created. The original patient type variable was a combination of the patient class (public or private) and the source where the patient was being referred from. These variables were split into two, as they were two different variables that have the potential to add value to the analysis. The appointment date and time were split into two variables, one for the date and another for the time. From the date variable, the day of the week for the appointment was calculated and the time of the appointment was categorised into a morning or afternoon appointment. The waiting time in days was also added as a new

variable and is the difference between the order date and the appointment date. In the final column 0 was added to identify the patient as attended. Table 3.5 provides a snippet of the attended dataset.

Table 3.5 Snippet of Attended dataset

pseudo mrn	appointment requested on	Appointment date	Time	Appointment day	Time of day	Waiting time in days
8925588	29/01/2018	29/01/2018	09:45	Mon	0	0

Referral Source	Modality	Age	Gender	Patient Class	Order Priority	Did Not Attend
Outpatient	Magnetic Resonance Imaging	76	M	PUBLIC	Routine	0

Table 3.6 New Variable list of the Attended dataset

Variable name	Description
Pseudorandom MRN	Patient Unique Identifier
Appointment requested on	Date appointment requested
Appointment date	Appointment date
Appointment time	Appointment time
Appointment day	Appointment day
Time of day	Morning or Afternoon appointment

Waiting time	Wait time from date appointment requested to scheduled date
Referral source	Source of referral
Modality	Type of imaging scan
Age	Age
Gender	Gender
Patient class	Public or private
Order priority	Prioritised appointment status
Did not attend	Appointment attended or missed

3.3.6 Merging of Both Data Sets

To have a complete file for analysis both datasets were merged. There were 18 fields in the DNA data set altogether with 14 fields in the attended dataset after processing. Fields from the DNA data set were removed to match column headers to convert the data into the same format of fields. Appointment month was removed as there was insufficient attended data to allow analysis of the month factor; as the attended data set included only the months of January and February 2018. Morning and afternoon fields were both merged into one column assigning 0 for morning and 1 for afternoon to match the attended data. The distance field was also removed as I had no access to patient addresses in the attended dataset. The scheduled status field was also removed as this had no bearing in the analysis therefore was irrelevant. The final list of fields before any descriptive statistics were analysed are represented in Table 3.6. After merging both data sets, the set of standardised, and provisionally considered suitable data for logistic regression analysis was 8,431 appointments (attended and did not attend) before descriptive statistics were analysed. The attended data,

was from January to February 2018 and the DNA set was from November 2014 to December 2018.

3.4 Preliminary Descriptive Statistics of Frequencies

Preliminary descriptive Statistics of the data ready for logistic regression was to gain more insight into the dataset with the possibility of removing potential noise that may skew the results of the regression analysis with the focus of using the best cases for analysis.

3.4.1 Waiting time in days Variable

After running frequency statistics for each of the variables to be used in regression analysis, the waiting time in days variable showed an imbalance in the number of waiting time in days (see figure 3.1): many scans (18% or 1552) were ordered and carried out on the same day (waitingtimeindays = 0). Of these appointments, fewer than 1% were marked as DNA. It seems these are appointments for patients admitted to hospital and scanned on the same day; these are unlikely to miss the scan. Including these appointments in the data set will skew the analysis and so they were removed.

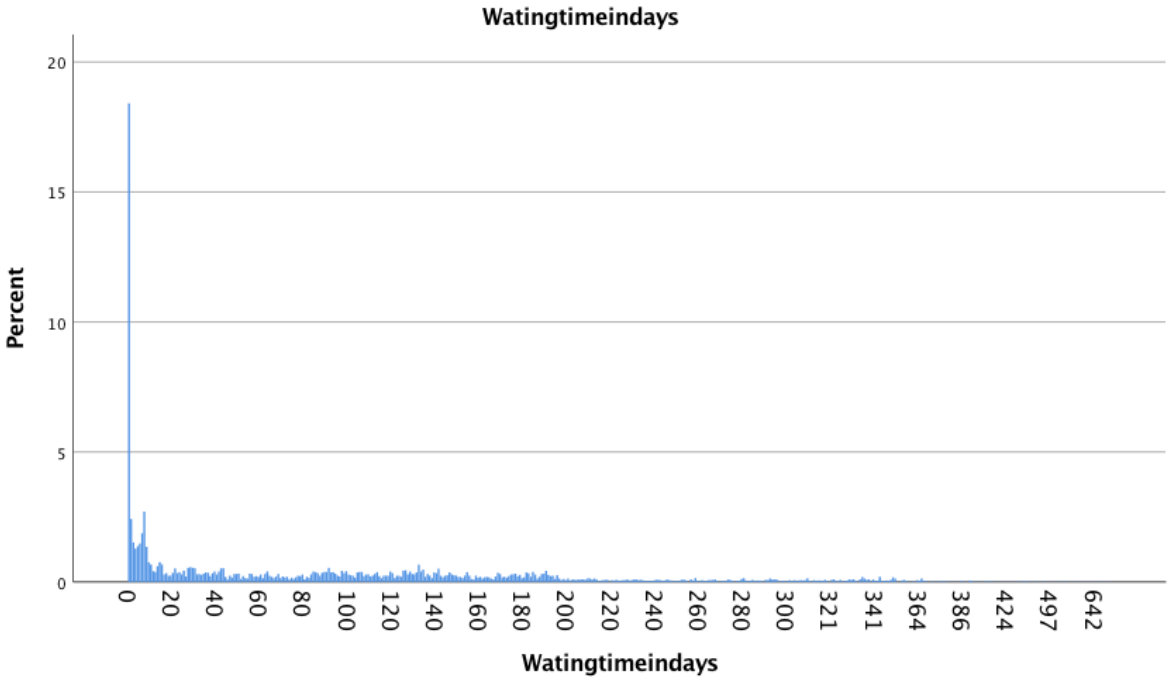


Fig. 3.1 Waiting time in Days Variable

3.4.2 Days of the week Variable

The days of the week variable was also studied after the waiting time in days was cleaned. This uncovered more imbalances in the data. The dataset included appointments from Monday to Sunday (see Table 3.7). Saturday and Sunday appointments were found to have an imbalance between the attended and DNA. There were unexplained errors on those days. As a result of the appointment entries that had 0 waiting days to the scheduled appointment that were removed from the data in the previous section imbalances were uncovered for Saturday and Sundays. As shown in table 3.7, for Saturdays, there was only 2 appointments marked as attended and 75 as DNA, and for Sundays there was no attended and 102 DNA. An explanation to these imbalances are appointments for emergency scans with patients admitted on Friday and Saturday that may have got their scan done on the day. According to (Hosmer DW, Lemeshow S, 2013) as a rule of thumb, there should be a minimum number of 10 observations per independent variable in a data set. Therefore, Saturday and Sunday appointments were removed and excluded from the analysis.

Table 3.7 Days of the week Variable

	Appointment day						
	Mon	Tue	Wed	Thu	Fri	Sat	Sun
Attended	908	950	971	918	1,021	2	
Did not attend	326	393	448	440	325	75	102

The Dataset now included 6,700 appointments (Table 3.8)

Table 3.8 Number of Overall Appointments

Attendance					
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Attended	4768	71.2	71.2	71.2
	Did not Attend	1932	28.8	28.8	100.0
	Total	6700	100.0	100.0	

3.5 Data Protection

All the data supporting this study was provided and approved by an Irish hospital. Given the immense sensitivity of the patient and health related information obtained measures were taken to anonymise the data, as described above. The data protection guidelines on research in the health sector were followed. Figure 4.2 presents best practice approaches to undertaking research projects using personal data as outlined by the data protection commissioner in Ireland.

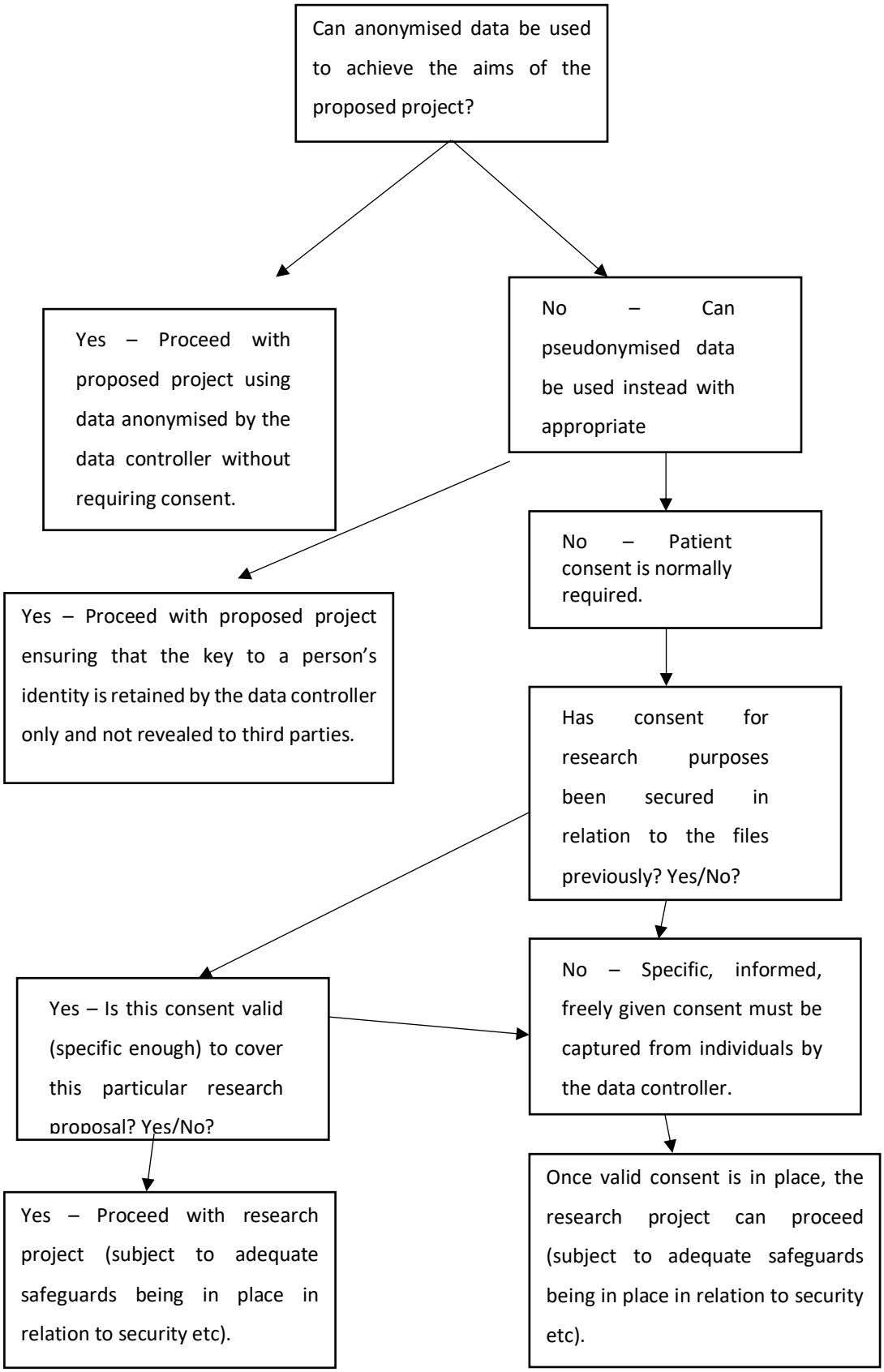


Fig. 3.2 Data protection Guidelines on research within the health sector

The data I received was already partially anonymised by the hospital. Patient medical record number was converted into a pseudorandom number and the date of birth was converted to age before the data was obtained. Furthermore, the addresses were included in the DNA data set after it was obtained and I converted addresses to distance from hospital. I was the only person to have access to the data.

3.6 Description of the data

3.6.1 Attendance

The combined dataset used to build a prediction model, consisted of 6,700 appointments, of these were 4,768 were recorded as attended over the period of January and February 2018 and 1,932 were recorded as DNA for the period of November 2014 to December 2017

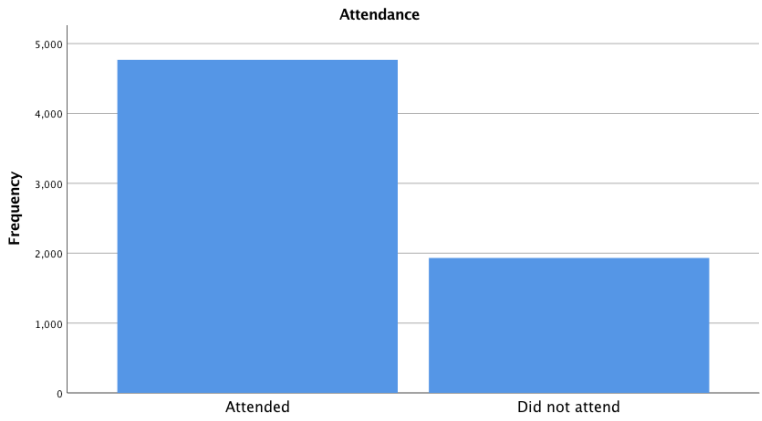


Fig. 3.3 Attended vs Did not Attend

Table 3.9 Attended vs Did not Attend (Figures)

Attendance					
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Attended	4768	71.2	71.2	71.2
	Did not attend	1932	28.8	28.8	100.0
	Total	6700	100.0	100.0	

3.6.2 Morning and Afternoon appointments

57.2% of appointments were in the morning and 42.8% attended in the afternoon. Appointments from 8am to 12pm were considered as morning appointments and afternoon appointment were after 12pm (Table 3.10)

Table 3.10 Morning and Afternoon appointments (Figures)

Timeofday					
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Morning	3833	57.2	57.2	57.2
	Afternoon	2867	42.8	42.8	100.0
	Total	6700	100.0	100.0	

3.6.3 Days of the week

There were appointments for all 5 days of the week. The day with the minimum number of appointments was Monday and the busiest day of the week was a Wednesday (Table 3.11 and Fig. 3.4).

Table 3.11 Days of the week (Figures)

Appointmentday					
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	MON	1234	18.4	18.4	18.4
	TUE	1343	20.0	20.0	38.5
	WED	1419	21.2	21.2	59.6
	THU	1358	20.3	20.3	79.9
	FRI	1346	20.1	20.1	100.0
	Total	6700	100.0	100.0	

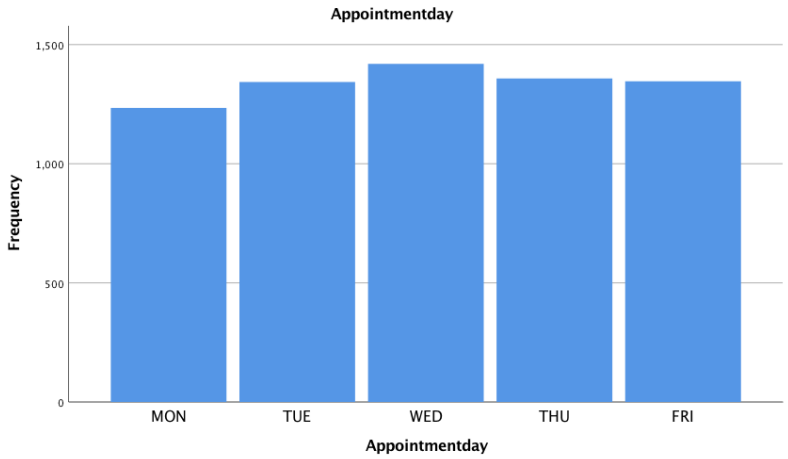


Fig. 3.4 Days of the week

3.6.4 Referral Source

The referral source column is where the patient was referred to the radiology department from. There are 4 referral source values, of which 1 is assigned to the appointment. Day case is when a patient has visited the hospital for a day procedure and has been sent for a radiology scan. GP referrals are from the general practitioner a patient has visited at some stage and they have been referred for scans. The Outpatient value is when a patient has visited the hospital for a consultation and was sent for a scan after. The Referral value is when a patient has been sent to the hospital from an external source such as another hospital or clinic (Table 3.12 and Fig. 3.5)

Table 3.12 Referral Source (Figures)

ReferralSource		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Daycase	169	2.5	2.5	2.5
	GP	740	11.0	11.0	13.6
	Outpatient	4220	63.0	63.0	76.6
	Referral	1571	23.4	23.4	100.0
	Total	6700	100.0	100.0	

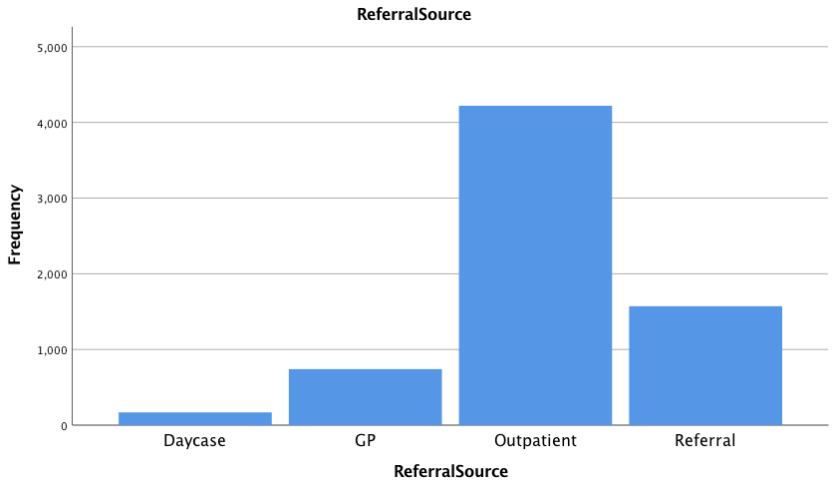


Fig. 3.5 Referral Source

3.6.5 Modality

The modality column describes the type of scan the appointment is for. They are categorised into 4 modalities: Computed Tomography, Magnetic Resonance Imaging, Mammogram and Ultrasound. 47.1% of appointments have been made for an ultrasound scan, with mammography in second with 20% (Table 3.13 and Fig. 3.6).

Table 3.13 Modality (Figures)

Modality					
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Computed Tomography	1168	17.4	17.4	17.4
	Magnetic Resonance Imaging	1038	15.5	15.5	32.9
	Mammography	1341	20.0	20.0	52.9
	Ultrasound	3153	47.1	47.1	100.0
	Total	6700	100.0	100.0	

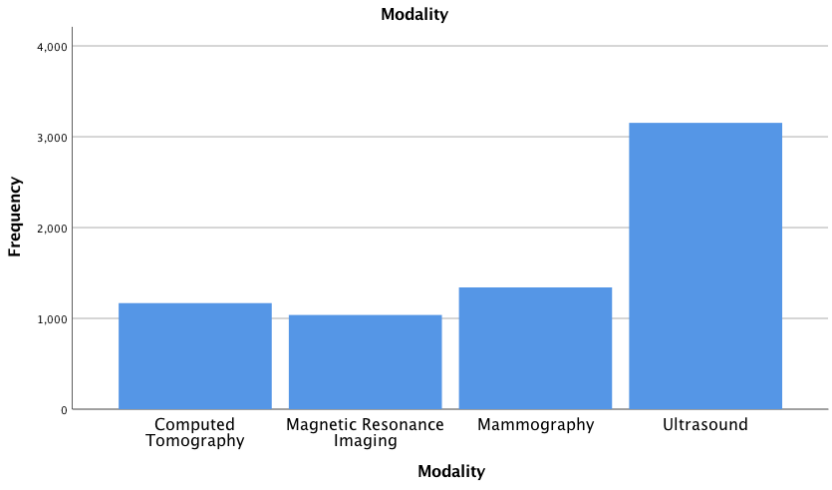


Fig. 3.6 Modality

3.6.6 Males vs Females

71.3% of appointments were for females and 28.7% were for males (Table 3.14 and Fig. 3.7).

Table 3.14 Gender (Figures)

Gender					
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	F	4777	71.3	71.3	71.3
	M	1923	28.7	28.7	100.0
	Total	6700	100.0	100.0	

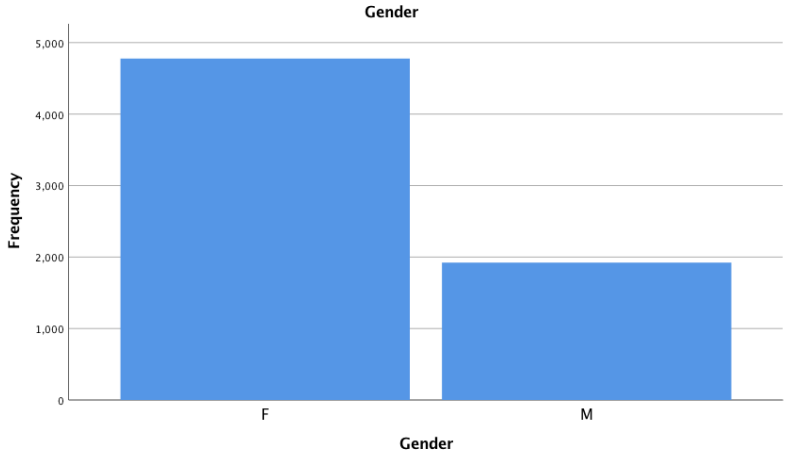


Fig. 3.7 Gender

3.6.7 Patient Class

The Patient class variable describes a two-tier sector, identified as public or private sectors. Patients that are considered as Private have healthcare services provided by entities other

than the government. Public sector healthcare is provided and covered by the government. 87.6% of appointments were public sector (Table 3.15 and Fig. 3.8).

Table 3.15 Patient class (Figures)

PatientClass		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	PRIVATE	831	12.4	12.4	12.4
	PUBLIC	5869	87.6	87.6	100.0
	Total	6700	100.0	100.0	

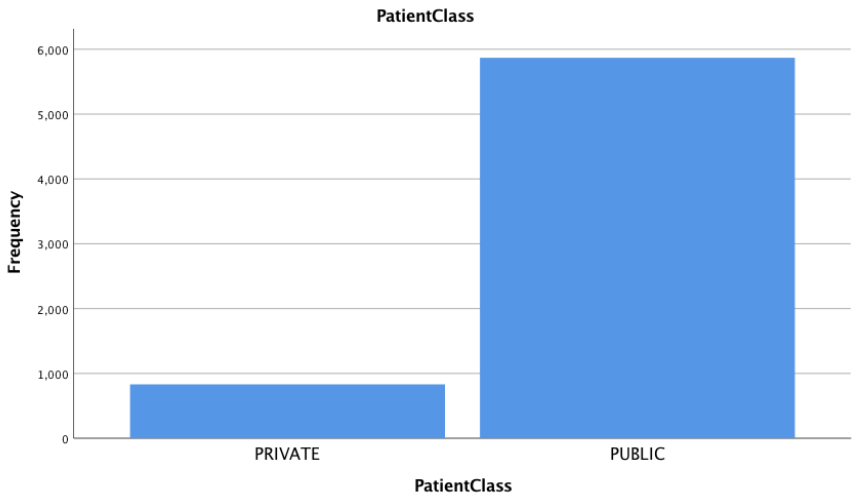


Fig. 3.8 Patient Class

3.6.8 Order Priority

Order priority is a field that describes the type of appointment. There are different types of Order priorities: Clinical trial, Planned-follow up, Routine check-up and Urgent. As patients go through the various stages of scans and treatments, planned-follow up appointments are scheduled to assess the current condition of the patient possibly in the phase of treatment. Routine priority appointments can be scans every few years possibly after the patient has been treated and recovered and can be part of regular procedures with no specific reason. Urgent is a top priority scan in which an appointment is made as soon as possible with little waiting time. In the data set, there is a small number of appointments for patients who are participating in clinical trials and they have been included in the analysis (Table 3.16 and Fig. 3.9).

Table 3.16 Order priority (Figures)

OrderPriority		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Clinical Trial	19	.3	.3	.3
	Planned Follow Up	669	10.0	10.0	10.3
	Routine	4420	66.0	66.0	76.2
	Urgent	1592	23.8	23.8	100.0
	Total	6700	100.0	100.0	

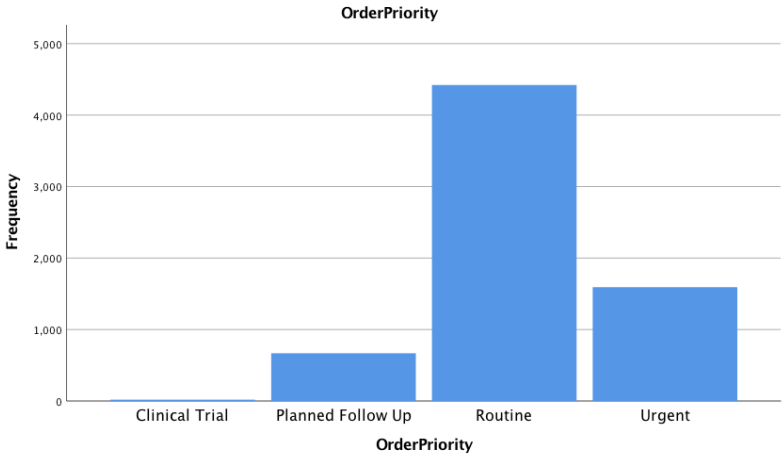


Fig. 3.9 Order Priority

3.6.9 Waiting Time in Days

Waiting time in days was calculated by taking the order request date and calculating the number of days the patient had to wait for the appointment. Appointments with 0 waiting days have been removed. Now this ranges from 1 to 755 days, with an average wait time of 109.82 days. The median mark is at 92 days and the mode is at 7 days (Table 3.17 and Fig. 3.10).

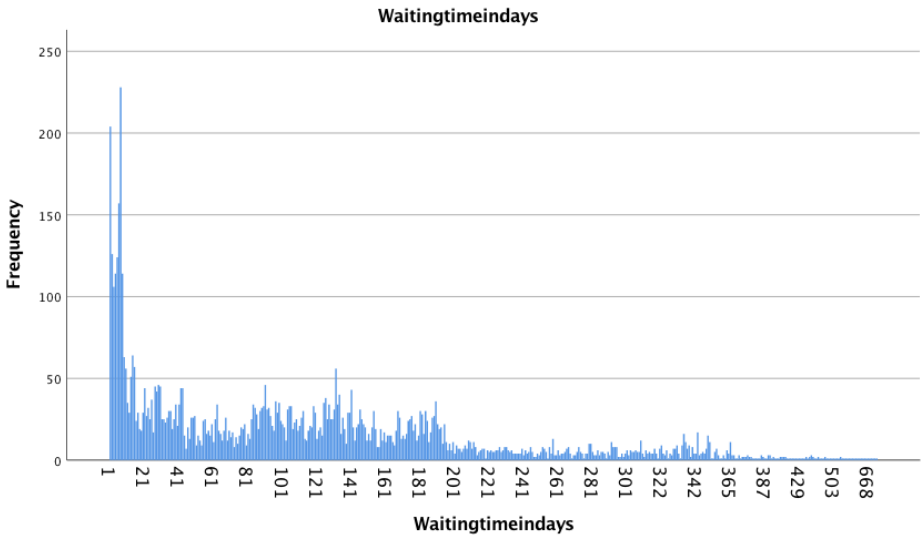


Fig. 3.10 Waiting time in Days

3.6.10 Age

The age of patients ranges from 7 to 98 years old (Table 3.17 and Fig. 3.11). There are more appointments (171 or 2.6%) for 47 year olds than any other age. The average age is 52.46 years old.

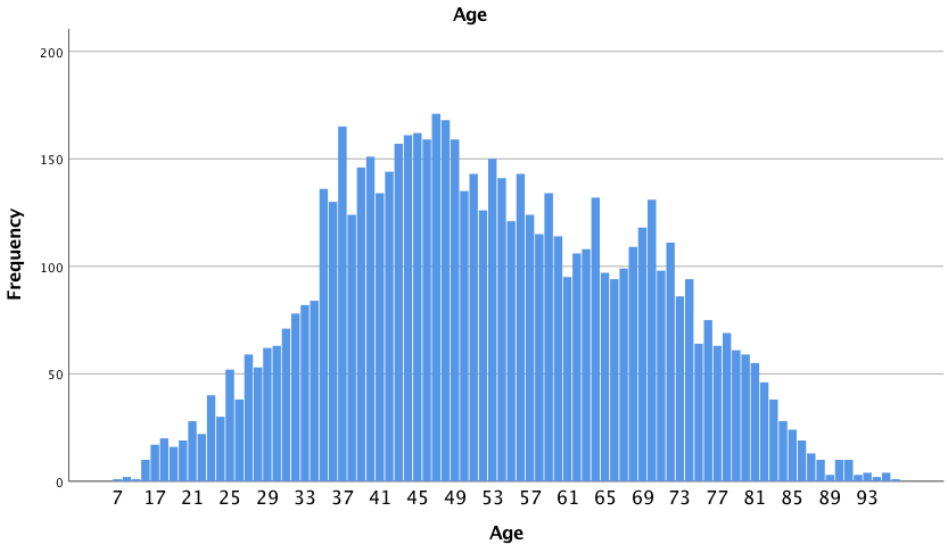


Fig. 3.11 Age

Table 3.17 Central Tendencies (Figures)

Central Tendencies			
		Waitingtimeindays	Age
N	Valid	6700	6700
	Missing	0	0
Mean		109.82	52.46
Median		92.00	51.00
Mode		7	47

3.7.1 Attendance variable with respect to numerous variables in the dataset

The attendance variable was compared against numerous other variables to understand the type of patient that attended or missed an appointment. This contains a description of the dataset rather than an analysis of the dataset. The most important factors in this study are the attended and DNA, which was how the pairs were chosen to study the type of patients that attend and miss appointments. These will be the two main factors compared to other factors that are included in the dataset. Variables are compared and graphed against each other. All figures and percentages are based from the combined data set of attended and DNA appointments.

3.7.2 Attendance vs Gender and Modality

Out of 6,700 appointments 4 modalities were studied and their attendance rates with respect to gender. Appointments for Ultrasounds were highest, with a combined gender DNA rate for ultrasounds at 15.85% of the overall dataset. The lowest DNA rates was for CT scans for both genders with a DNA rate of 3.03%. There was also a small DNA rate for males attending a mammogram scan at 0.10% (Table 3.18 and Fig. 3.12).

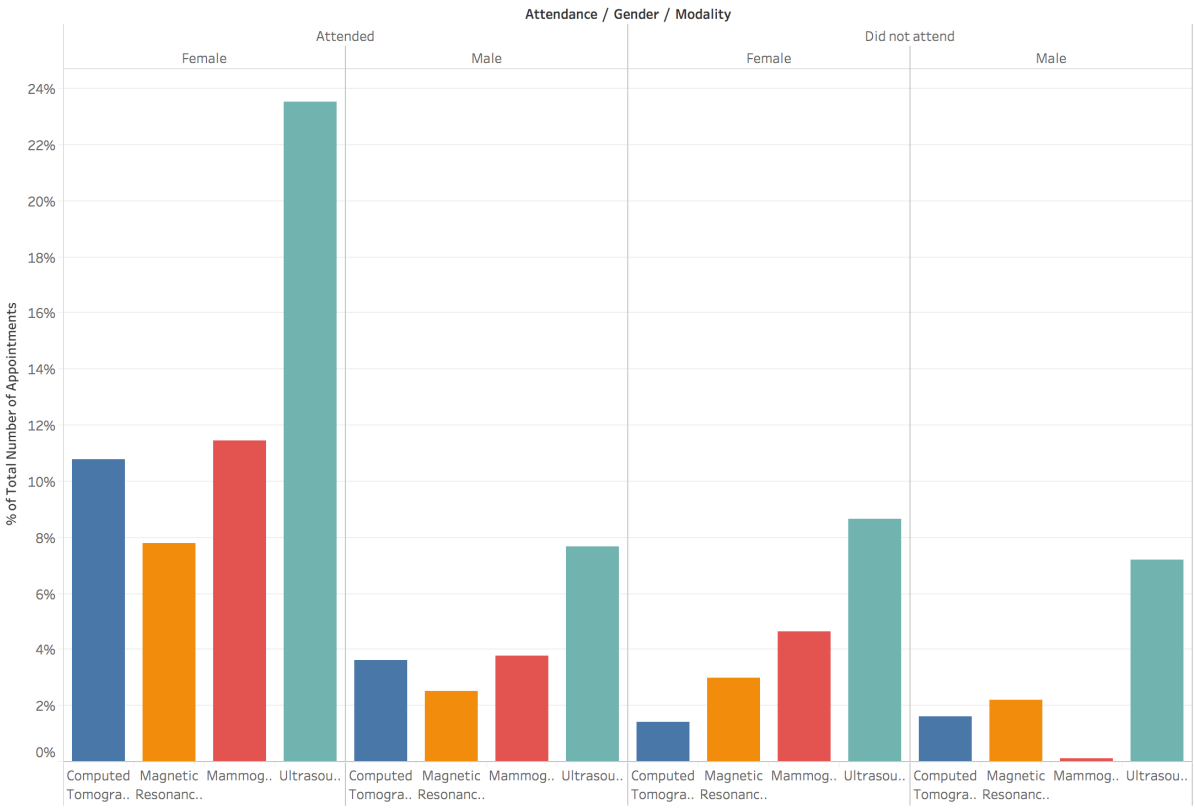


Fig. 3.12 Attendance vs Gender vs Modality

Table 3.18 Attendance vs Gender vs Modality (Figures)

		Gender	
Attendance	**Modality**	**Female**	**Male**
Attended	Computed Tomography	10.79%	3.61%
	Magnetic Resonance Imaging	7.78%	2.52%
	Mammography	11.46%	3.79%
	Ultrasound	23.54%	7.67%
Did not attend	Computed Tomography	1.42%	1.61%
	Magnetic Resonance Imaging	3.00%	2.19%
	Mammography	4.66%	0.10%
	Ultrasound	8.66%	7.19%

3.7.3 Attendance vs Patient class and referral source

Looking at the attendance for public and private healthcare patients for the different referrals, lowest rate of non-attendance were for patients with private healthcare, with a day case referral. Day cases in hospitals are when patients are undergoing a procedure and possibly sent for a scan to analyse the success of the procedure. Since the patient maybe be in the hospital anyway this could explain the low DNA rate for day cases. Fewer than 1% of day case appointments were DNAs. However, highest non-attendance rates were within the outpatients referral source with public patients, at 20.6% of the overall dataset (Table 3.19 and Fig. 3.13).

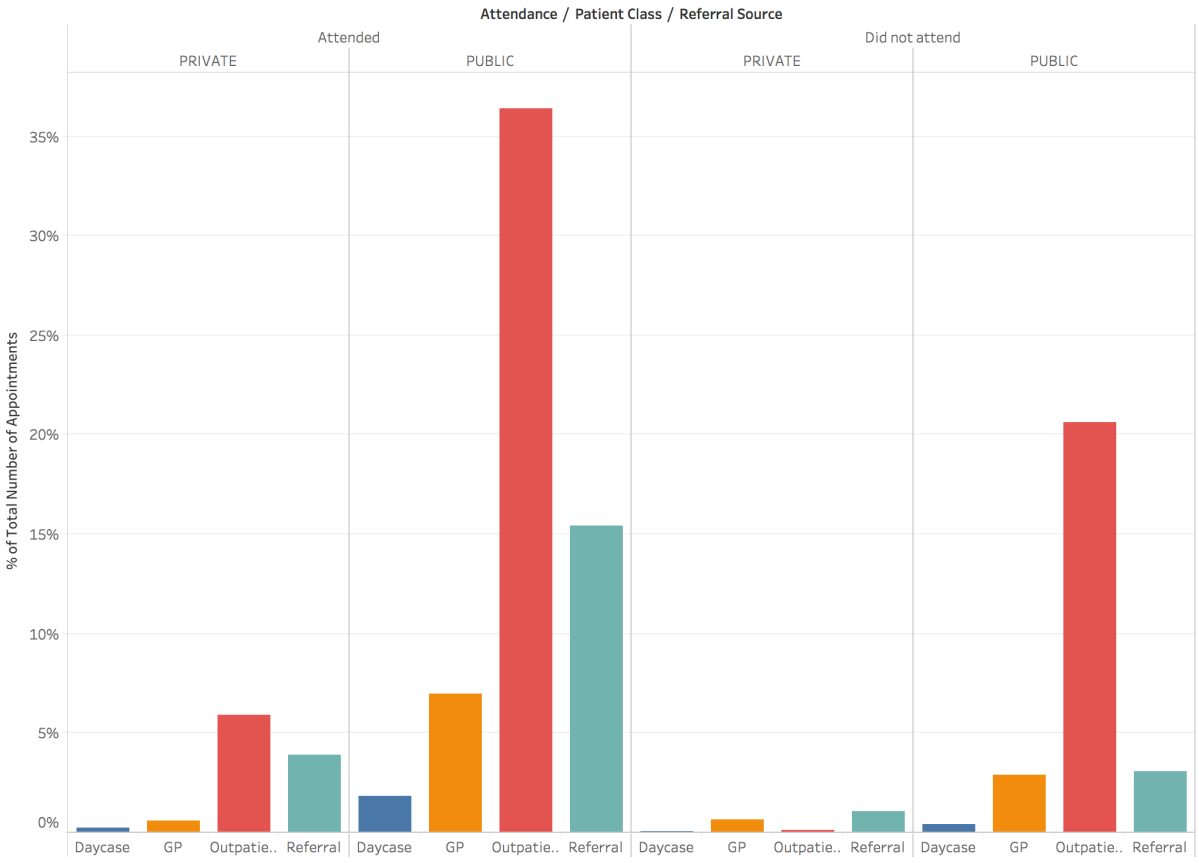


Fig. 3.13 Attendance vs Patient class vs Referral Source

Table 3.19 Attendance vs Patient class vs Referral Source (Figures)

		Patient Class	
Attendance	Referral Source	PRIVATE	PUBLIC
Attended	Daycase	0.21%	1.84%
	GP	0.57%	6.97%
	Outpatient	5.90%	36.40%
	Referral	3.88%	15.40%
Did not attend	Daycase	0.06%	0.42%
	GP	0.63%	2.88%
	Outpatient	0.09%	20.60%
	Referral	1.07%	3.09%

3.7.4 Attendance vs Time of day and day of the week

Appointments were divided into morning and afternoon appointments by using the appointment time. Factoring in the day of the week, we can see that highest non-attendance rates by day of the week and morning or afternoon. Wednesday morning was the highest number of non-attendance in patients at 4.22% of all morning appointments. Thursday afternoon appointments saw the highest non-attendance rates at 2.77% of all afternoon appointments. Overall, the day of the week with the most missed morning and afternoon appointments was Wednesday with 6.68% of patients missing appointments (Table 3.20 and Fig. 3.14).

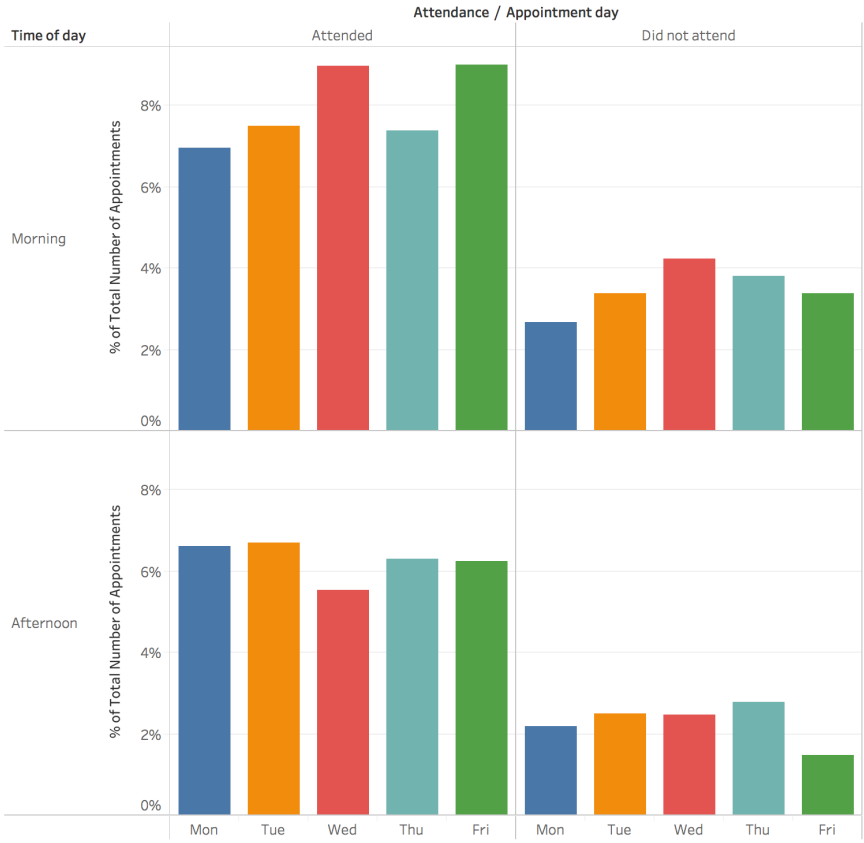


Fig. 3.14 Attendance vs Time of day and day of the week

Table 3.20 Attendance vs Time of day and day of the week (Figures)

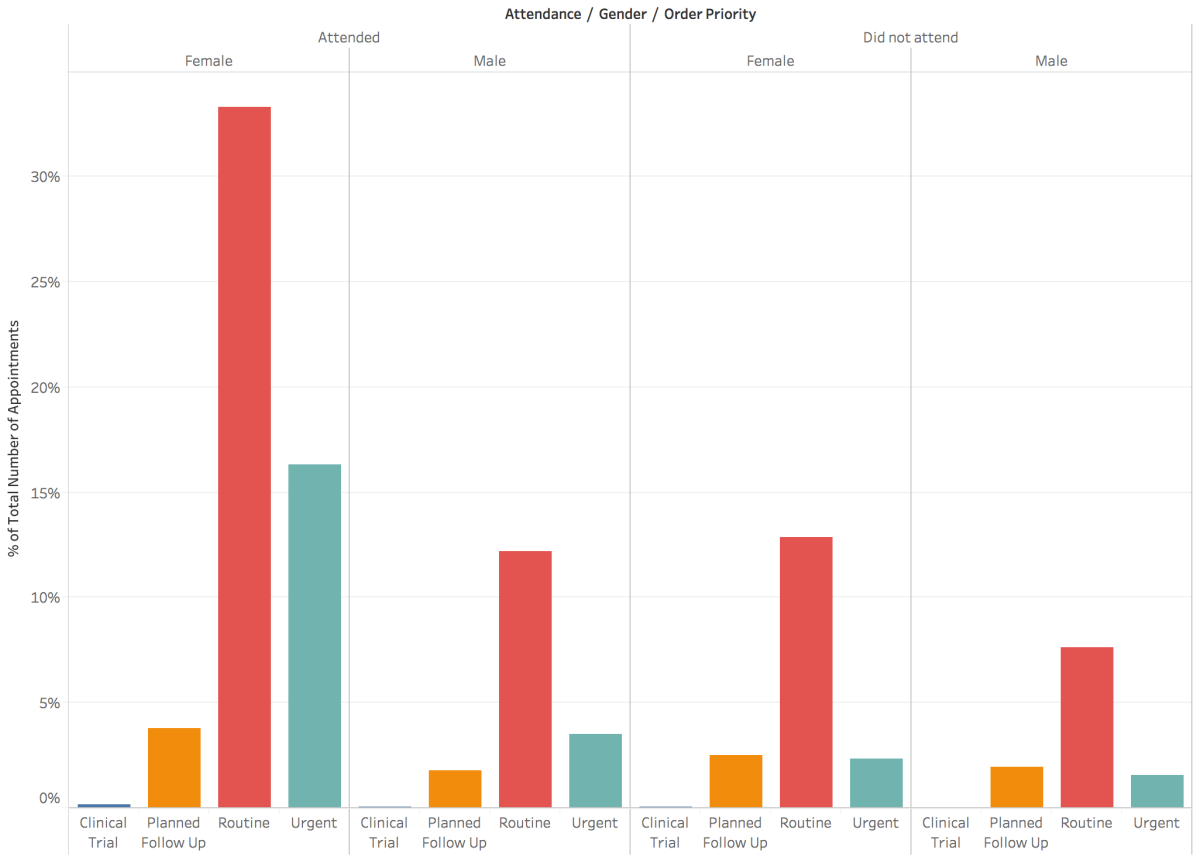
		Time of day	
Attendance	Appointment day	Morning	Afternoon
Attended	Mon	6.940%	6.612%
	Tue	7.493%	6.687%
	Wed	8.955%	5.537%
	Thu	7.388%	6.313%
	Fri	9.000%	6.239%
Did not attend	Mon	2.672%	2.194%
	Tue	3.373%	2.493%
	Wed	4.224%	2.463%
	Thu	3.791%	2.776%
	Fri	3.373%	1.478%

3.7.5 Attendance vs Gender and Order Priority

Looking at the order priority of the appointments, we have 4 sets. Clinical trial, Planned follow-up, routine and urgent. Looking at these categories against attendance and gender

shows us that highest DNA rates are in routine check-up appointments for females, which makes routine check-ups the largest order priority in the data set. Routine female attended and routine male attended appointments gives us respectively 33.31% and 12.18% of the overall data. As for the non-attendance in routine female or male comes to 12.85% and 7.63%. For every 11 routine check-ups, female patients there was 4 attended male patients and for every 13 routine check-ups, female patients that did not attend there was 8 males that did not attend. Highest non-attendances in Males with routine planned check-ups or observed in the dataset (Table 3.21 and Fig. 3.15).

Table 3.21 Attendance vs Gender vs Order Priority



		Gender	
Attendance	Order Priority	Female	Male
Attended	Clinical Trial	0.15%	0.07%
	Planned Follow Up	3.76%	1.81%
	Routine	33.31%	12.18%
	Urgent	16.34%	3.52%
Did not attend	Clinical Trial	0.04%	0.01%
	Planned Follow Up	2.49%	1.93%
	Routine	12.85%	7.63%
	Urgent	2.34%	1.55%

Fig. 3.15 Attendance vs Gender vs Order Priority (Figures)

3.8 Conclusion of descriptive statistics

This chapter provided a summary of the descriptive statistics of the dataset that will be included in the analysis. The goal of descriptive statistics is to describe in numerical format with the aid of charts to show what is currently happening within a known population. By understanding this information and graphing it, makes it easier to understand the statistical analysis to be conducted on this dataset. This chapter is the foundation for the predictive model to be developed in the next chapter.

Chapter 4 Development of the predictive model

4.1 Introduction

The purpose of this study is to develop a predictive model that predicts patients at high risk of not attending an appointment with the use business intelligence tools. Business intelligence software is a type of application software that is used to retrieve, analyse, transform and report data. SPSS was the statistical software used to develop the predictive model using binary logistic regression. Using this model probabilities are calculated for attended and missed appointments and an ROC curve generated to evaluate the quality of the model. The process in which the model is generated is presented in this chapter.

4.2 Development of the predictive model

The proposed model for predicting patients at high risk of not attending an appointment, is constructed using Binary logistic regression. Binary Logistic regression deals with how predictor variables are selected and entered into the model. Like all regression analysis, logistic regression is predictive analysis. It is used to describe the data and explain the relationship between one dependent binary variable and one or more nominal, ordinal or interval variables. In our case our dichotomous dependent variable is attended or did not attend, 0 or 1 respectively. The model will be used to analyse patient appointment data obtained from an Irish hospital imaging department.

Binary logistic regression analysis was conducted for the merged datasets of the attended and did not attend dataset of 6,700 appointments included in the study. These were the chosen data sets that reflected the full set of appointments. The analysis was run several times. The first run of the analysis was conducted as a preliminary run to analyse the results and asses the performance of the model. The second run was to remove variables with odd results that have a potential to throw off the overall model. In the second run, the predicted probability values were calculated and saved as a variable in the data set. The predicted probability values range from 0 to 1 for every appointment and are defined as

the probability that a particular outcome is a case divided by the probability that it is a non-case. Once these values were generated a new column was created in the dataset and a value assigned to every appointment. Using the predicted probability value, a receiver operating characteristic curve is generated to illustrate the diagnostic capability of the binary logistic regression model to discriminate attendance rates. The model was then used 3 different times with different cut-off points to generate the classification for various cut-off points. The flow of the modelling process is presented in Fig. 4.1

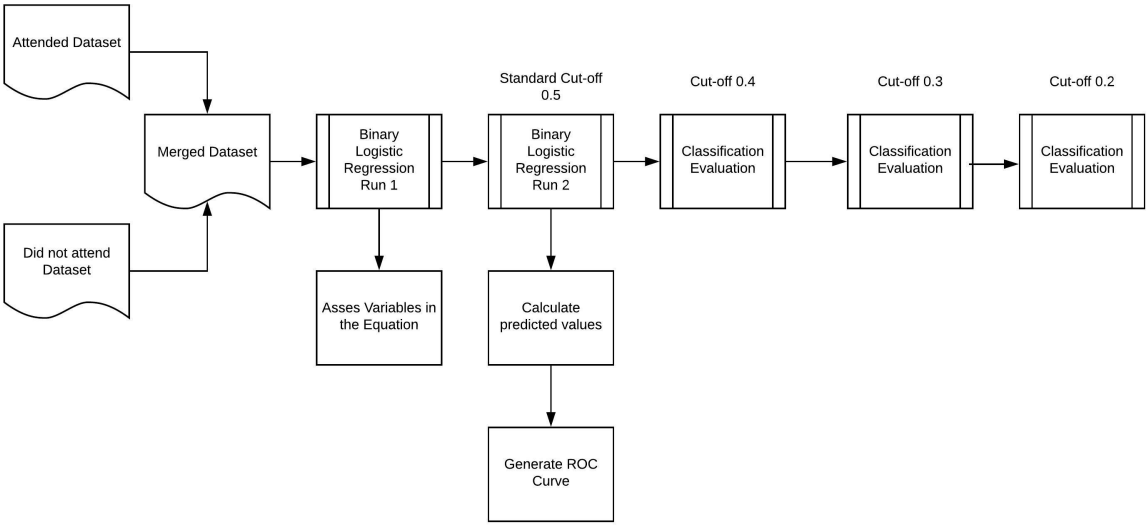


Fig. 4.1 Flow of Modelling Process

4.3.1 Output of Binary Logistic Regression Test run 1

The first run of the binary logistic regression model produces results presented in Table 4.5 with a standard cut-off of 0.5. We can see 6,700 cases have been included in the test regression analysis (Table 4.1). The Hosmer and Lemeshow goodness of fit was also calculated (Table 4.4).

Table 4.1 Case processing summary

		N	Percent
Selected Cases	Included in Analysis	6700	100.0
	Missing Cases	0	.0
	Total	6700	100.0
Unselected Cases		0	.0
Total		6700	100.0

The omnibus test of model coefficients of the regression analysis shows that the model was a poor fit for the prediction of appointment non-attendance (X^2 (4) = 943.56, p < 0.001) (Table 4.2). The effect size ranges from 13% to 18% (see Table 4.3) which is measured using the pseudo R squares (Cox and Snell, and Nagelkerke). According to Cohens effect size this is small in terms of prediction and classification accuracy with an outcome of 73.7% as seen in the classification table (Table 4.5). The result of the Hosmer and Lemeshow test suggests that there is a significant difference between actual and predicted values (X^2 (8) = 30.05, p < 0 .01 (see Table 4.4). The p <0.001 indicates that our data does not fit the model for non-attendance. Ideally, if the p > 0.05 this would indicate the data fits the model.

Table 4.2 Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	943.559	18	.000
	Block	943.559	18	.000
	Model	943.559	18	.000

Table 4.3 Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	7107.293	.131	.188

Table 4.4 Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	30.050	8	.000

The overall percentage accuracy classification of the model is 73.7%. The sensitivity of the classification of DNA's is 27.2% .525 cases were correctly predicted to be as DNA. The specificity of the cases that had the observed characteristic (Attended) was at 92.6% which is

4,416 cases were correctly predicted to be attended. The Positive Predicted Value (PPV) is 59.93% and the Negative Predicted Value (NPV) is at 75.82%.

Table 4.5 Classification Table

			Predicted		
			Attendance		
	Observed		0	1	Percentage Correct
Step 1	Attendance	Attended - 0	4416	351	92.6
		Did not attend - 1	1408	525	27.2
	Overall Percentage				73.7

4.3.2 Variables in the equation test run 1

The variables included in the test run of the regression analysis are presented in table 4.6. SPSS automatically reduced the degrees of freedom in some variables due to redundancies. The referral source, order priority, modality and day of the week variable are all missing one category. Referral, urgent, ultrasound and Wednesday were removed from the model. In the binary logistic regression output window, SPSS produced a warning; *Due to redundancies, degrees of freedom have been reduced for one or more variables*. This is a sign that these variables are linearly dependent. They can be inverses or sums of other variables. For example, the referral source had four categories of which only three were included. Accounting for these linearly dependent variables would have not affected the results.

We can see that the $p > 0.05$ for the order priority of clinical trials and the appointment day of Tuesday and Thursday. This shows that there is weak evidence that clinical trials, Tuesday and Thursday play a factor in appointment non-attendance. These 3 factors stand out above

all in the test run, their p values are abnormally high and suggest the data has a large standard error or the number of outcomes in the groups are small, but as they are not anywhere near the significance value they will be removed in the next run. However, there was evidence in the remaining variables with $p < 0.01$ or suggesting patient non-attendance is associated with these variables (Table 4.6). A more detailed explanation of variables will follow up in the final run of the model.

Table 4.6 Variables in the Equation

		df	Sig.	Exp(B)	B
Step 1	**Waiting time in days**	1	.000	1.003	.003
	Referral Source	3	.000		
	Daycase	1	.001	2.072	.728
	GP	1	.000	2.732	1.005
	Outpatients	1	.000	3.126	1.140
	Modality	3	.000		
	Computed Tomography	1	.000	.302	-1.197
	Magnetic Resonance Imaging	1	.001	.767	-.266
	Mammogram	1	.000	.452	-.794
	Age	1	.003	.994	-.006
	Gender	1	.000	.572	-.559
	Patient Class	1	.000	.447	-.806

	Order Priority	3	.000		
	Clinical Trial	1	.392	1.668	.512
	Planned Follow up	1	.000	4.273	1.452
	Routine	1	.000	2.217	.796
	Time of day	1	.001	1.225	.203
	Appointment day	4	.000		
	Monday	1	.001	.731	-.378
	Tuesday	1	.178	.886	-.314
	Thursday	1	.555	1.054	.052
	Friday	1	.000	.685	-.121
	Constant	1	.000	.182	-1.701

4.4.1 Output of Binary Logistic Regression run 2

The Second run of the binary logistic regression model produces the classification accuracy table presented in Table 4.11. In this run the checked box to calculate predicted probabilities was set with the standard cut off value of 0.5. We can see 6,681 cases have been included in the regression analysis (Table 4.7). The Hosmer and Lemeshow goodness of fit was also conducted. From the test run that was conducted in the previous section we saw that Clinical trials category and the appointment day of Tuesday and Thursday had high insignificance p values. Appointments for clinical trials were removed and the appointment day variable was removed from every appointment. As seen in the previous run their p values were abnormally high and suggest the data has a large standard error or the number of outcomes in the groups are small. This drops the number of appointments in the dataset to 6,681 which were all included in the analysis of the second run.

Table 4.7 Run 2 Case processing Summary

		N	Percent
Selected Cases	Included in Analysis	6681	100.0
	Missing Cases	0	.0
	Total	6681	100.0
Unselected Cases		0	.0
Total		6681	100.0

The omnibus test of model coefficients of the regression analysis shows that the model again was a poor fit for the prediction of appointment non-attendance (X^2 (4) = 913.3, $p < 0.001$) (Table 4.8). The effect size ranges again from 13% to 18% which is measured using the pseudo R squares (Cox and Snell, and Nagelkerke). According to Cohens effect size this is small in

terms of prediction and classification accuracy with an outcome of 74.1% as seen in the classification table (Table 4.11). The result of the Hosmer and Lemeshow test suggests that there is a significant difference between actual and predicted values (X^2 (8) = 33.81, p < 0 .01 (Table 4.10). The p <0.001 again indicates that our data does not fit the model for non-attendance. Ideally, if the p >.05 would indicate the data fits the model.

Table 4.8 Run 2 Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	913.304	13	.000
	Block	913.304	13	.000
	Model	913.304	13	.000

Table 4.9 Run 2 Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	7117.390[a]	.128	.183

Table 4.10 Run 2 Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	33.881	8	.000

As shown in Table 4.11 the overall percentage accuracy classification of the model is at 74.1% with the standard cut-off of 0.5, which is a slight improvement from the first run which had a

percentage accuracy classification of 73.7%. The sensitivity of the classification of DNA's is 27.4%. 528 cases were correctly predicted to be as DNA. The specificity of the cases that had the observed characteristic (Attended) was at 93.1%% which is 4,422 cases were correctly predicted to be attended. The Positive predicted value (PPV) is at 61.5% and the negative predicted value (NPV) is at 75.9%.

Table 4.11 Run 2 Classification Accuracy

			Predicted		
			Attendance		
	Observed		0	1	Percentage Correct
Step 1	Attendance	0	4422	330	93.1
		1	1401	528	27.4
	Overall Percentage				74.1

4.4.2 Variables in the equation run 2

The variables included in the test run of the regression analysis are presented in table 4.12. SPSS automatically reduced the degrees of freedom in some variables due to redundancies. Referral Source, Modality and order priority are all missing one factor each. Referral from the referral source variable, Ultrasound from the modality variable and urgent from the order priority variable have been removed due to linearity. Looking further into the p values of the variables we can see that all of them are statistically significant with $p < 0.01$.

Table 4.12 Run 2 Variables in the Equation

		df	Sig.	Exp(B)	B
Step 1	**Time of day**	1	.002	.826	-0.191
	Waiting time in days	1	.000	1.003	0.003
	Referral Source	3	.000		
	Day Case	1	.001	2.120	0.751
	GP	1	.000	2.725	1.002
	Outpatients	1	.000	3.056	1.117
	Modality	3	.000		
	Computed Tomography	1	.000	.294	-1.223
	Magnetic Resonance Imaging	1	.002	.769	-0.262
	Mammogram	1	.000	.451	-0.796
	Age	1	.005	.995	-0.005
	Gender	1	.000	.568	-0.565
	Patient Class	1	.000	.447	-0.806
	Order Priority	2	.000		
	Planned Follow-up	1	.000	4.242	1.445
	Routine	1	.000	2.230	0.802
	Constant	1	.000	.192	-1.651

4.5.1 Calculating Threshold – ROC Curve

In the second run of the regression analysis the sensitivity and specificity was calculated at 27.4% and 93.1%. To assess the discrimination of the model an ROC curve was generated. The predicted values were also generated in a newly created column. For the ROC Curve all possibilities of the cut-off points are considered and plotted. The ROC curve was generated by entering the predicted values into the test variable box and the state variable was the attendance variable set to a state of 1 for did not attend. The AUC is presented in Fig. 4.2.

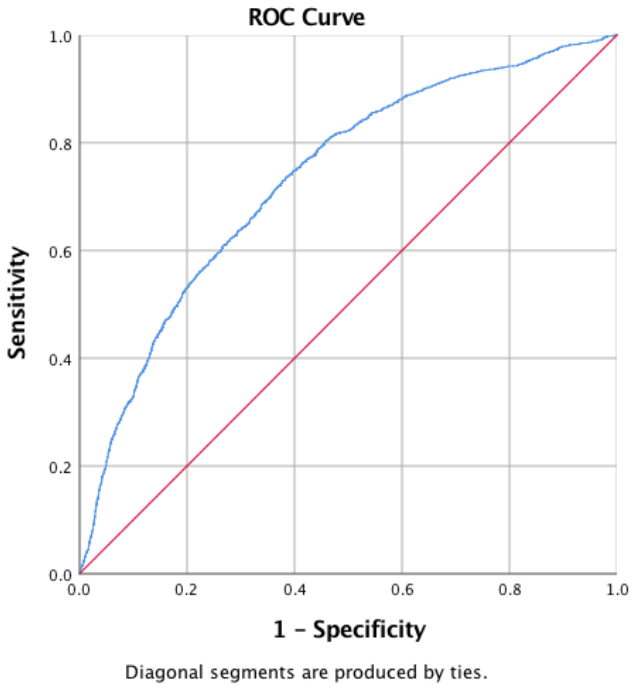


Fig. 4.2 ROC curve

4.5.2 Area under the curve

We can see that the AUC is at 0.734 in table 4.14. The area can range from 0.5 to 1, with higher values representing better discrimination. According to (Hosmer and Lemeshow, 2000), a value of 0.734 puts this at an acceptable level of discrimination with a 95% confidence interval from 0.721 and 0.747 (Table 4.14). The AUC curve has given us an acceptable level of discrimination but the Hosmer and Lemeshow test has stated that the model is a poor and inadequate fit. (Kramer and Zimmerman, 2007) assessed the validity of the Hosmer and Lemeshow test for various sample sizes and concluded that a significant Hosmer and Lemeshow test doesn't necessarily mean that a predictive model is not useful, and proved that it may be inaccurate with larger sample sizes starting from 5000. The general rule of thumb as outlined by (Hosmer and Lemeshow, 2000) in Table 4.13 is:

Table 4.13 (Hosmer and Lemeshow, 2000)

AUC	**Classification**
AUC = 0.5	This suggests no discrimination, so we might as well flip a coin
0.7 ≤ AUC < 0.8	We consider this acceptable Discrimination
0.8 ≤ AUC < 0.9	We consider this excellent Discrimination
AUC ≥ 0.9	We consider this outstanding Discrimination

Table 4.14 Area Under the Curve

Area	Std. Error[a]	Asymptotic Sig.[b]	Asymptotic 95% Confidence Interval	
			Lower Bound	Upper Bound
.734	.007	.000	.721	.747

4.6 Classification Evaluation with various cut-off points

The data was classified using the second model and the various cutt-off points of 0.2, 0.3, 0.4 and 0.5 were calculated. As shown in Fig. 4.1 the flow of the modelling process included the standard classification cut-off of 0.5 as a standard. Using this Binary logistic regression run, different cut-off points were evaluated. Additionally cut-off points of 0.2, 0.3, 0.4 were assessed. The sensitivity, specicity, PPV and NPV were calculated for all cut off points in table 4.25. The Classification model will be explained in more detail in the next chapter.

Table 4.15 Classification Evaluation

Classification Cut-off	**0.2**	**0.3**	**0.4**	**0.5**
Sensitivity	86.3%	67.3%	47.3%	27.4%
Specificity	43.4%	66.8%	83.6%	93.1%
Positive predicted Value	38.24%	45.10%	53.93%	61.53%
Negative predicted value	88.65%	83.40%	79.61%	75.94%

4.7 Conclusion of the development of the predictive model

One of the aims of this dissertation is to develop a model that can predict patients at high risk of not attending an appointment. The model overall has been developed but with several runs of trial and error. Variables to be included in the run were assessed and chosen carefully to avoid skewing the results. A test run was conducted to assess the variables in which some were removed. The conclusion and result of this chapter is a model that predicts patients at high risk of not attending using Binary logistic regression.

Chapter 5 Interpretation and application of the predictive model

5.1 Introduction

In this chapter the results of the predictive model will be presented and interpreted from the tables produced by SPSS. How the results address the research question will be discussed followed by the significance of the results and the limitations of the study.

5.2.1 Results of the model

The results of the second run of Binary logistic regression analysis was applied on 6,681 patient appointments. Since the first run was a test run, the second run concluded the B coefficients, Exp(B) and the p values that will be used to form the equation further in the chapter. The binary classification model dependent variable was attended vs did not attend, respectively 0 and 1, respectively. The model revealed that all variables included and their relevant categories were statistically significant with $p < 0.01$. The variables in the equation from chapter 4 are revisited and discussed in more detail.

Table 5.1 Variables in the Equation Results

		df	Sig.	Exp(B)	B
Step 1[a]	**Time of day**	1	.002	.826	-0.191
	Waiting time in days	1	.000	1.003	0.003
	Referral Source	3	.000		
	Day Case	1	.001	2.120	0.751
	GP	1	.000	2.725	1.002
	Outpatients	1	.000	3.056	1.117
	Modality	3	.000		
	Computed Tomography	1	.000	.294	-1.223
	Magnetic Resonance Imaging	1	.002	.769	-0.262
	Mammogram	1	.000	.451	-0.796
	Age	1	.005	.995	-0.005
	Gender	1	.000	.568	-0.565
	Patient Class	1	.000	.447	-0.806
	Order Priority	2	.000		
	Planned Follow-up	1	.000	4.242	1.445
	Routine	1	.000	2.230	0.802
	Constant	1	.000	.192	-1.651

5.2.2 Order priority

The order priority variable consisted of planned follow-up, routine and urgent visits types. These were automatically recoded by SPSS into dummy variables. These are variables created by SPSS to represent the subgroups of the variable that can be recognized as a category so regression analysis can be performed. Each category (dummy variable) in this case had 0 coded for other order priorities and 1 for the specific order priority. The categories of planned follow-up and routine were highly significant compared to all three order priorities with $p < 0.01$. The odds ratio for planned follow up appointments is significantly higher with patients and 4.2 times more likely not to attend a planned follow up appointment compared to a routine or urgent appointment, and 2.2 times more likely not to attend a routine appointment compared to a planned follow up or urgent appointment. We have no results for the urgent category in the order priority variable as the category was dropped by SPSS due to redundancies. Degrees of freedom have been reduced for this variable and made no difference to the overall results.

5.2.3 Modality

The modality variable included Computed Tomography, Magnetic Resonance Imaging and Mammogram categories. These were automatically recoded by SPSS into dummy variables. These are variables created by SPSS to represent the subgroups of the variable that can be recognized as a category so regression analysis can be performed. Each category (dummy variable) in this case had 0 coded for other modalities and 1 for the specific modality. Computed tomography, Mammograms and Magnetic resonance imaging all had $p < 0.01$. The odds ratio for computed tomography indicated that there is decreased odds of 0.294 for an increase in one unit of computed tomography (0 for other modalities and 1 for CT). Since computed tomography was coded as a dummy variable by SPSS the inverted interpretation of this would be, for each unit reduction (0 for other modalities) patients are 3.4 times more likely to miss other modality appointments which include mammogram, ultrasound and magnetic resonance imaging compared to computed tomography appointments. Compared with Magnetic resonance imaging and mammogram the odds ratios are 0.769 and 0.451. Both

modalities indicate decreased odds with every unit increase. These ratios inverted give 1.3 and 2.2 respectively with the odds of non-attendance increasing by those factors.

5.2.4 Referral source

The referral source variable included Day case, GP and outpatients categories. These categories are where the patient has been referred from. Day Case is when a patient has been in for a procedure and has been refereed for a scan. GP referral source is when a GP may have some concerns and has referred the patient for a scan. Outpatients referral sources are patients visiting the outpatients departments for various reasons and have been referred to the imaging department for a scan. These factors also contributed to patient non-attendance. GP and outpatients both have a significance value $p < 0.01$. Patients referred from the outpatients department are over 3 times more likely to miss an appointment compared to day case and GP referrals.

5.2.5 Time of day

The time of day considered morning and afternoon appointments. 0 was assigned to morning appointments and 1 to afternoon appointments. The model revealed that time of day was statistically significant with a $p < 0.01$. With every unit increase (from morning to afternoon) in the time of day variables a patient has 0.8 chances of missing an appointment which inversely states that patients are 1.25 times more likely to miss a morning appointment compared to afternoon appointments.

5.2.6 Waiting time in days

Waiting time in days was highly significant with a $p < 0.01$. The odds ratio stated that with every unit increase in days waiting a patient is 1.003 times more likely to miss an appointment.

5.2.7 Age

The model revealed that the younger a patients is, the more likely they are to miss appointments. With a statistical significance value $p < 0.01$ the odds ratio stated that with increase in age patients were less likely to miss appointments. Inversely stated with every unit decrease in age a patient is 1.005 times more likely to miss an appointment.

5.2.8 Gender

Males were coded as 0 and females were coded as 1 in the model. Gender was highly significant stating that with a unit increase in gender (i.e. female) they were 0.5 times less likely to miss an appointment, suggesting that males are 1.7 times more likely to miss appointments.

5.2.9 Patient class

Public and private patients attending appointments were assigned 0 and 1 respectively in the model. This variable was also highly significant suggesting that public patients are over 2 times more likely to miss a scheduled appointment than private patients

5.3 ROC curve analysis

The AUC gave us 0.734, which is an acceptable level of discrimination. The cut-off point of 0.5 gave us a sensitivity of 27.2% and a specificity of 92.6%. When the cut-off point was increased to 0.734 according to the ROC curve this decreased the sensitivity to 0.9% and increased the specificity to 99.7%, lowering the overall percentage classification accuracy by 2.5%. The AUC states that any predicted value above 0.7 will give us an acceptable level of discrimination in the model.

5.4 Hosmer and Lemeshow test

In addition, the Hosmer and Lemeshow test was used to test for the goodness of fit of the model. It assesses if the observed event rates match the expected event rates in the subgroups. This was the option in SPSS that was chosen. The subjects were divided in to deciles based on their probabilities predicted and the chi-square was calculated from observed and expected recurrences. This shows that the model was a poor and inadequate goodness of fit with a $p < 0.01$, therefore statistically significant. Ideally if the $p > 0.05$, the model would have been a good fit. As mentioned previously (Kramer and Zimmerman, 2007) assessed the validity of the Hosmer and Lemeshow test for various sample sizes and concluded that a significant Hosmer and Lemeshow test doesn't necessarily mean that a predictive model is not useful, and proved that it may be inaccurate with larger sample sizes starting from 5000. The Hosmer and Lemeshow test has been a de facto standard in many software packages. (Hosmer *et al.*, 1997) even acknowledged that there are several drawbacks with using this test and the most troubling problem is that results can depend on the number of deciles. So we cannot rely on this test for the fit of the model.

5.5 Practical Application of the Predictive model

Consider a 76 year-old male patient attending an afternoon appointment in the outpatients department as a public patient for a routine MRI scan with a wait time of 254 days to the appointment.

To demonstrate the application of this model for the above appointment, using table 4.12 and applying the logistic regression equation, let p be the probability of attendance with the model generalized as the following:

$$\log\left[\frac{p}{1-p}\right] = \alpha + (\beta_1 X_1) + (\beta_2 X_2) + (\beta_3 X_3) + (\beta_4 X_4) + (\beta_5 X_5) + (\beta_6 X_6) + (\beta_7 X_7) + (\beta_8 X_8)$$

The above equation applied to the above patient gives us:

$$\log\left[\frac{p}{1-p}\right] = \alpha + (\beta_1 * Time\ of\ day) + (\beta_2 * Waiting\ time\ in\ days) + (\beta_3 * Referral\ source) + (\beta_4 * Modality) + (\beta_5 * Age) + (\beta_6 * Gender) + (\beta_7 * Patient\ class) + (\beta_8 * Order\ priority)$$

The input of the co-efficient for example is Time of day is β_1, from the last column in table 4.12 and the constant is considered as α. From the sample appointment, the patient has an appointment in the afternoon ($\beta_1 = -0.191$) and ($X_1 = 1$). With 254 days to the appointment ($\beta_2 X_2 = 0.762$). With a referral source from the outpatients clinic ($\beta_3 X_3 = 1.117$); for an MRI scan ($\beta_4 X_4 = -0.262$) who is a 76 years old ($\beta_5 X_5 = -0.38$) male ($\beta_6 X_6 = 0$) (males coded as 0 and females as 1), attending as a public patient ($\beta_7 X_7 = 0$) (public class patients coded as 0 and private as 1) for a routine scan ($\beta_8 X_8 = -0.802$).

The calculation of the probability of attendance is:

$$\log\left[\frac{p}{1-p}\right] = -1.1651 + (-0.191) + (0.762) + (1.117) + (-0.262) + (-0.38) + (0) + (0) + (0) + (0.802) = 0.197$$

Therefore,

$$\frac{p}{1-p} = e^{0.197} = 1.217745$$

$$p = \frac{1.217745}{1 + 1.217745}$$

$$p = 0.549$$

According to the model the above patient has a probability of 54.9% of not attending the scheduled appointment.

A sample appointment was chosen from the dataset because of the high predicted probability to create this scenario. The predicted probability is nearest to the maximum which was 0.83 predicted by the model contains the following characteristics, A 54-year-old female who has been waiting for the appointment for 504 days has been referred from the outpatients clinic as a public patient for an MRI scan for a planned follow up visit with a morning appointment scheduled. This patient has a predicted probability of 0.83, a probability closer to 1 is a patient likely to miss an appointment.

Another sample appointment taken from the dataset which was chosen because of the low predicted probability value to create this scenario. The patient is highly likely to attend with a probability of 0.01 which was predicted by the model contains the following characteristics, A 31-year-old female who has been waiting for an appointment for 74 days has been referred from another hospital as a private patient for a CT scan for a routine check-up with an afternoon appointment scheduled. The patient had a predicted probability of 0.01. A probability closer to 0 is a patient that will highly likely attend the appointment.

5.6 Classification Evaluation

In table 4.25 the classification evaluation of the predictive model was presented, with cut-off points 0.2, 0.3, 0.4, 0.5. Since the Hosmer and Lemeshow test concluded that the model was a poor fit and proved that there was a large difference between actual and predicted values, the main aim was to minimize this. Every modelling method may have errors and these are distinguished into two categories. For this case type 1 errors means reporting a DNA as an attended and type 2 errors means reporting an attended patient as a DNA. We can see from the classification table 4.11 with a cut-off point of 0.5 that the Type 1 error is 72.6% and the Type 2 error is 6.9%. A cut-off point was chosen by maximising the sum of sensitivity and specificity. According to (Cantor *et al.*, 1999), the cut-off determination by maximizing the sensitivity and specificity is the equivalent to choosing a point in the ROC curve when the slope of the tangent is 1. This method minimizes the sum of the false negatives and false positives misclassification likelihoods. In table 5.2 the cut-off point that utilizes this is 0.3 with a sum of 134.1% (Specificity + Sensitivity). The sensitivity and the specificity both added for the three cut-off points 0.2, 0.3 and 0.4 are all very close, 129.7%, 134.1% and 130.9% respectively. Perhaps the somewhere between 0.2 and 0.3 as well as 0.3 and 0.4, there maybe be a higher maximization point more than 134.1%

Table 5.2 Maximising the Sum of Sensitivity and Specificity

Classification Cut-off	0.2	0.3	0.4	0.5
Sensitivity	86.3%	67.3%	47.3%	27.4%
Specificity	43.4%	66.8%	83.6%	93.1%
Sensitivity + Specificity	129.7%	134.1%	130.9%	120.5%

The ROC curve gave us an acceptable level of discrimination to identify patients at high risk of not attending an appointment and by maximizing the sum of sensitivity and specificity we were able to calculate a second possible cut-off point of 0.3. Considering we are trying to calculate the patients at high risk of not attending an appointments we have room for error taking into consideration the target intervention methods that the hospital may have in place. Unlike other application of classification models that are based on epidemiological studies the cost of an error is much higher if classifying a patient as not having the disease and the disease being present compared to classifying a patient as attending and they do not attend.

As shown in Table 5.3 the overall percentage accuracy classification of the model is at 66.9% with a cut-off of 0.3. The sensitivity of the classification of DNA's is 67.3%. 1298 cases were correctly predicted to be as DNA. The specificity of the cases that had the observed characteristic (Attended) was at 66.8%% which is 3,172 cases were correctly predicted to be attended.

Table 5.3 Classification table for cut-off point 0.3

			Predicted		
			Attendance		
	Observed		0	1	Percentage Correct
Step 1	Attendance	0	3172	1580	66.8
		1	631	1298	67.3
	Overall Percentage				66.9

5.7.1 Target intervention to reduce DNA's

The use of medical diagnostic imaging devices is becoming more wide spread as technology advances. Missed appointments at radiology departments are specifically very costly to clinics and hospitals due to waste of time on expensive imaging equipment and specially trained staff that need to operate the equipment. As estimated by the HSE the average cost of a missed appointment in Irish health care was put at €44 per missed appointment. These costs arise from reviewing referral notes, communication back and forth between GP's and specialists and obtaining medical records. For the nuclear medicine department, costs can be much higher when factoring in radiopharmaceuticals. Considered a medical supply, some radioactive materials have a two-hour half-life and so its degradation can be costly when a patient doesn't show.

Using business intelligence tools and the specific area of predictive analytics, this can help in reducing many costs within a hospital, decrease lost revenue while focusing in patient-

centeredness. When discussing business intelligence tools and predictive analytics for the benefit of the patient it is important to ask how the statistics affect the end users knowledge. With the use of statistical analysis there are many opportunities to facilitate the shift of raw data to knowledgeable and useful data that can improve the quality of care for patients and reduce costs (Sibte, Abidi and Yusoff, 1998). Decisions can be made that according to the end users knowledge, ways of resolving consistent issues that may arise. Missed appointments is a consistent issue within every hospital or clinic. Predictive analytics can lead to a strong decision support resource for hospital management if used in the correct way, by effectively using data already recorded from EHR and scheduling systems that is readily available to most hospitals or clinics. This topic of decision making and acting on the knowledge the end user has gained from predictive analytics introduces various method of target intervention for missed appointments.

5.7.2.1 Communication

Different methods of communication are used within many hospitals for scheduling, reminders and informing patients of results. The different methods of communication have been by postal letters, email, text messaging and phone calls. SMS which stands for short message service is probably the most popular and widely used data application in the world. It is often referred to as texting and is a 2-way alphanumeric method for sending messages across a network whether it be from mobile phone to mobile phone or web to mobile phone. Sending texts is normally limited to 160 characters and are often split into several messages when this limit is reached.

5.7.2.2 SMS application to missed appointments

Application of the predictive model generated in the previous chapter can be applied to the different communication methods discussed. Much of the literature online covers SMS texting for appointment reminders and very little is discussed regarding older methods of communication such as postal letters. At a cut-off rate of 0.3 the model gives us a sensitivity of 67.3% for patients who are most likely not to attend an appointment. These patients should

opt into reminders when their details are registered. However, all patients should be asked to opt into reminders but the select few that are highly likely not to attend should be the primary focus. Since literature reviews stated previously by (Junod Perron *et al.*, 2013) there was no difference in DNA rates in comparison to telephone reminders and SMS text reminders, and since SMS text reminders are the cheaper alternative, SMS should be considered. A bi-directional SMS system should be setup where the patient receives a confirmation SMS text of the appointment asking the patient to reply with a yes or a no. A second friendly reminder should be sent on the day before the appointment scheduled date. At a cost of €0.14 for each patient this should be one of the most cost effective methods for reminders, considering the cost of a missed appointment in Irish health care has been calculated at €44 by the HSE. Automated systems can be readily available to reduce the need for human interaction and cut costs to schedule the SMS text reminders and patients who reply with a no at the initial confirmation should be followed up by a telephone call to reschedule if needed. A cancellation is not ideal but far less costly than a missed appointment. This can open-up booking slots and gives hospital administration staff time to reschedule appointments to patients who need it most. Although research found has been in different European countries further investigation is needed from an Irish context to study the effectiveness of SMS reminders.

5.7.3.1 Overbooking

Overbooking approaches have also been suggested in literature reviews with the use predictive modelling. Overbooking appointments mitigates the lost productivity for missed appointments. Overbooking is defined by (LaGanga and Lawrence, 2015) by adjusting the time intervals between appointments, using block scheduling of multiple patients at one or more schedule times. Double booking is a form of overbooking and is a specific case of block-booking, which schedules a multiple number of patients to show up at the same time. In the case of the imaging department patients attending can be overbooked or an extreme case double-booked. For example, two patients attending a mammogram can be overbooked and shortening the time slot of the. Naïve overbooking is a common practice in many hospitals to

address patient missed appointments. It is considered naïve because it is implemented without any prior knowledge of the probability that a patient will attend. Naïve overbooking can lead to negative impacts for hospitals and clinics resulting in increased patient wait times as mentioned by (LaGanga and Lawrence, 2007).

5.7.3.2 Overbooking application to missed appointments

Overbooking approaches can be applied to the predictive model demonstrated in the previous chapter to allow for more efficient scheduling using the predicted probability calculated and the cut-off point of 0.3. Table 5.4 is an example of how we can apply the over booking approach to the predictive model using a 2-tier approach, one for booking appointments by overlapping appointments and shortening time slots (overbooking) and a more extreme case of booking two appointments at the same time (double-booking). Table 5.2 is an example of a day in a scheduling system of patients to attend with 30 minute slots for each appointment. Each appointment has a predicted probability of how likely that patient is to attend. Using the threshold of 0.3, patients with a probability of less than 0.3 is likely to attend and more than 0.3 is at risk of not attending by maximizing the sum of sensitivity and specificity. The overbook column contains a No for keeping the schedule as is, or Overbook or Double-book according to the predicted probability. As shown there are two patients with a probability of 0.46998 and 0.56418 that have been calculated by the predictive model. These are likely not to attend therefore that time schedule should be over booked. Following the ROC curve we calculated the AUC at 0.734. This is the cut-off point for discrimination between attended and did not attend and we can apply this to the threshold. We can use the AUC as an extreme case where the patient is very likely not to show to an appointment by maximising the sensitivity. Using 0.734 from the ROC curve gives us a 95% confidence interval from 0.721 and 0.747. If a patient has a calculated predicted probability of over 0.734 we can be 95% confident that a patient will miss an appointment and therefore we can use the double-booking approach to minimise costs and physician time in a hospital. We have one patient with a predicted probability of 0.90040 who is a very high risk patient that we can double book his appointment. Table 5.4 shows 16 appointments selected from the data set with predicted probabilities specifically selected to demonstrate an example of this. The time

column was kept to 30 minute slots. A new time schedule can be created from this with overbooking and double booking implemented.

Table 5.4 Schedule Booking Example

Time	Wait time in days	Referral source	Modality	Age	Gender	Patient class	Order Priority	Predicted Probability	Overbook or Double-book
8:00	1	Referral	Magnetic resonance imaging	55	Female	Public	Routine	0.12306	No
8:30	87	Outpatient	Magnetic resonance imaging	46	Female	Public	Urgent	0.20885	No
9:00	141	Outpatient	Ultrasound	50	Female	Public	Routine	0.46998	Overbook
9:30	6	Referral	Magnetic resonance imaging	29	Female	Public	Routine	0.14058	No
10:00	6	Outpatient	Magnetic resonance imaging	54	Female	Public	Urgent	0.16417	No
10:30	43	GP	Computed Tomography	53	Female	Public	Routine	0.07028	No
11:00	190	Referral	Ultrasound	36	Female	Public	Routine	0.14040	No
11:30	9	Outpatient	Mammogram	58	Female	Public	Routine	0.22554	No
12:00	504	Outpatient	Ultrasound	54	Male	Public	Planned-Follow up	0.90040	Double-book
12:30	43	GP	Ultrasound	57	Female	Public	Routine	0.14165	No

13:00	317	Outpatient	Magnetic resonance imaging	33	Male	Public	Planned-Follow up	0.56418	Overbook
13:30	3	Referral	Ultrasound	65	Male	Public	Urgent	0.07240	No
14:00	2	Referral	Ultrasound	61	Female	Private	Urgent	0.07335	No
14:30	43	GP	Computed Tomography	37	Female	Public	Routine	0.07603	No
15:00	7	Outpatient	Ultrasound	66	Female	Public	Routine	0.14915	No
15:30	6	Referral	Ultrasound	37	Male	Public	Routine	0.16927	No

5.6 Variables under patient control

The overall number of variables that were included in this predictive model was eight. Out of these variables not many are under the control of the patient. Application of online web based appointment booking systems and have been proved to be beneficial, by giving the patient flexibility in choosing their own time slots. In the predictive model that was developed we may have the time of days to the scheduled appointment under the patients control to a certain extent and fully being able to control the morning or afternoon appointments. Statistically the morning and afternoon appointments show that patients are 1.25 times more likely to miss a morning appointments compared to afternoon appointments. For the wait time in days, every unit increase in days a patient is 1.003 times more likely to miss an appointment. Patients can choose the date whether it be several weeks or several months away, possibly avoiding summer months where a patient may be abroad on holidays. By implementing a choose and book system that allows a bi-directional input for the hospital and patients will lead to an increase in communication and flexibility for the patient therefore reducing patient non-attendance by allowing them to choose the most suitable date and time as opposed to the hospital sending out the schedule that the patient has no control over without back and forth communication from the hospital.

5.7 Limitations of the Study

Throughout the research there was several problems encountered that may have added to a less effective predictive model and inaccurate results.

1. There were many errors in the DNA dataset. This was an Excel file that was maintained manually by hospital administration staff. Spelling mistakes, internal codes and comment sections made standardising the data for logistic regression difficult and appointments with errors had to be removed. In comparison to the attended data set which was a direct export of the hospital scheduling system had better standards and was in a more readable format.

2. The inconsistencies between the DNA data set and Attended dataset were problematic. The DNA data set had more variables available for selection. The DNA data set that was obtained contained addresses that were anonymised and converted to distance from hospital that could not be used in the overall analysis. A considerable amount of work was done using google API's to convert these to distance from hospital. This was a very important factor in being able to predict patients travelling a further distance to attend appointments and could have considerably affected final results. In previous literature, this played an important role in results and was also a major assumption before any analysis was performed. The distance variable had to be removed as the attended dataset did not have this.

3. The attended data set included records only for the first two months of the year, January and February 2018, as opposed to the DNA dataset which included three years

of data. Months of the year could have been added to the regression model if the attended data set was for at least one year.

4. Days of the week were added but had to be removed as part of the first logistic regression run which was the test run to assess the model for the first time. The high statistical figures produced in the variables for the equation included in table 4.6 in the test run, for the appointment day showed an imbalance in the dataset, possibly between the modalities, referral source and order priority. For example, there may have been fewer than ten computed tomography appointments with a referral source from a GP with a planned follow-up order priority. As a rule of thumb for logistic regression the one in ten rule is applied for how many predictor variables derived from the data when doing regression analysis. This has been proven with the clinical trial order priority where there were only 19 appointments consisting of clinical trials but an imbalance between the attended and did not attend, therefore regression analysis could not be performed on these appointments. These appointments were removed from the analysis.

5. Patient history is also an important variable to consider that was not included in the model. Patients with previous missed appointments are also likely to continue this trend. In the attended data set there was no way of distinguishing returning patients from first time patients, whereas in the DNA dataset where Patient numbers were randomised but still retained on to the returning attribute.

5.8 Conclusion of the results and interpretation

Interpretation of the statistical analysis performed in the data was presented in plain English in this chapter. Each variable was assessed accordingly and the biggest factors contributing to missed appointments were presented. Limitations of the study were presented and this concluded that the model results could have been greatly improved.

Chapter 6 Conclusion

6.1 Introduction

This chapter outlines the key findings and outlines recommendations for future research and explains what can be done to improve the predictive model by comparing different models. The conclusion of the dissertation will be presented along with an individual reflection on the whole process. The aim of the study was to apply predictive analytic techniques with the concept of business intelligence to help predict patients at high risk of not attending an appointment.

6.2 Analysis and tools

The tools used for the analysis and predictive modelling were:

- Microsoft Excel 2016
- IBM SPSS Statistics 25
- Tableau Desktop 10.5

A student licence was obtained for the above tools

6.3 Key Findings

Historical data was used in this study to predict future non-attendances within the imaging department of a large Irish hospital. Business intelligence is a large technology driven process which answers questions about the data available and results in presenting actionable information to aid in decision making. Decisions made within the context of this study can

lead to better quality of care to patients, reduce wait times to appointments and dramatically reduce day to day operating costs. Predictive analytics techniques work, but only with the correct data processing methods and the predictive power of the model corresponds with how relevant the variables used in the model are and how accurate they maybe, in other words variable selection for the model is important. The key findings of this study are outlined below:

There are many factors that contribute to patient non-attendance within the radiology department. Results of the model concluded that there is not only one factor that affects patient non-attendance but a combination of factors all at once. Some of these factors can be manipulated by some of the target interventions mentioned and giving the patient the flexibility of choosing their appointments with the use of web based technologies.

- The model proved that factors from previous studies such as age, gender and waiting time in days all affect attendance rates.
- Patients tend to miss morning appointments rather than afternoon appointments
- Effective communication and overbooking techniques have been demonstrated to lower patient non-attendance rates

6.4 Recommendations for future research

This study was focused on the topic of predictive analytics using logistic regression and various other techniques. The implementation of the predictive model can be approached in different ways. Future work in this area can involve, obtaining more reliable and a larger data set and applying comparative statistical analysis such as decision trees. Logistic regression vs Decision trees models can be compared for a more effective model. Much of the research in this area consisted of quantitative methods for predicting patient non-attendance, many personal factors such as lateness, sickness forgetfulness and patient personal issues can be

assessed. It is difficult to predict such factors with patients but may consist of a qualitative study such as a survey that may add to the predictive model.

6.5 Contributions to the research

This study has focused and demonstrated the use of business intelligence tools and the application of predictive analytics within the Irish health care sector for predicting patients at high risk of not attending. A predictive model was developed, recommended cut-off points were calculated and proposed interventions. Although this study focuses within the area of radiology appointments, the predictive model and interventions in this study can be applied to any radiology department where appointments are scheduled. To my knowledge, I am not aware of any previous studies concerning patient attendance and predictive analytics conducted within an Irish healthcare context.

6.6 Individual Reflection

Health informatics is a multidisciplinary area which involves communications technology in health care, computer science, engineering and statistics. These are all areas which I have a deep interest in. With my background of software engineering my primary aim of this dissertation was to focus on the predictive analytics area of health informatics. This was a whole new concept for me, with no statistical or mathematical background I found this area of analysis very interesting. While reflecting on the experience of writing this dissertation I came to realisation that I have learned some invaluable methods on how predictive analytics work, how they can be applied to real life scenarios and has provided me the fundamentals of future work for my own personal development that I can pursue. I have also learned the value and power of Business intelligence tools. Tableau as a business intelligence and visualisation tool has been a slight challenge to learn throughout this dissertation and was used to create charts and tables in chapter 3. SPSS, the IBM analytics software has been an

enormous challenge to learn and apply logistic regression for this dissertation. I can now look back on this experience and realise this has helped me both as a student and as a professional. Research and academic writing skills are a valuable skill to learn from an academic and professional sense. Overall this has been an enjoyable and rewarding experience.

6.7 Conclusion

This dissertation highlights the importance of business intelligence concepts with the application of predictive analytics as a means of addressing patient non-attendance rates and various means of tackling these rates and leading to unprecedented intelligence on patient characteristics, needs and recognising cost saving opportunities. Research into business intelligence is a worthwhile investment and will play a significant role in the future of Irish hospital management. The first step that is highlighted is the importance of the data collected through various enterprise systems, in which there is enormous amounts of data available in healthcare. An important question that was considered throughout this dissertation is how can the statistical information from predictive analytics affect the end user's knowledge? To achieve this a predictive model was developed with various statistical methods to study the characteristics of patients that are at high risk of missing an appointment. Findings concluded that all variables associated with appointments and patients included in the model contributed to non-attendance. Given the quality and the quantity of the data obtained a more accurate model may have been achieved using more variables with a more accurate and complete dataset with all the attended and DNA appointments over a specific period greater than 1 year. Several cut-off points were calculated to aid in the discrimination of attended vs did not attend patients. Enhancing communication, overbooking appointment methods with prior predictive knowledge and allowing patient flexibility by being able to choose their own appointments are all successful methods of reducing non-attendance as proven in much of the previous literature and applied to the predictive model.

References

Alaeddini, A. *et al.* (2015) 'A hybrid prediction model for no-shows and cancellations of outpatient appointments', *IIE Transactions on Healthcare Systems Engineering*, 5, pp. 14–32. doi: 10.1080/19488300.2014.993006.

Atherton, H. *et al.* (2012) 'Email for the coordination of healthcare appointments and attendance reminders', *Cochrane Database of Systematic Reviews*. John Wiley & Sons, Ltd. doi: 10.1002/14651858.CD007981.pub2.

Bean and Andrew (1995) 'Predicting appointment breaking', *James Journal of Health Care Marketing; Spring*, 15(29). Available at: https://search.proquest.com/docview/232321231/fulltextPDF/354C3AFAD5D34B53PQ/2?accountid=14404 (Accessed: 6 January 2018).

Brannan, S. O. *et al.* (2011) 'The effect of short messaging service text on non-attendance in a general ophthalmology clinic', *Scottish Medical Journal*, 56(3), pp. 148–150. doi: 10.1258/smj.2011.011112.

Breiman, L. (2001) 'Random forests', *Machine Learning*, 45(1), pp. 5–32. doi: 10.1023/A:1010933404324.

Cantor, S. B. *et al.* (1999) 'A Comparison of C/B Ratios from Studies Using Receiver Operating Characteristic Curve Analysis'. Available at: https://ac-els-cdn-com.elib.tcd.ie/S089543569900075X/1-s2.0-S089543569900075X-main.pdf?_tid=169f4cfc-4737-48d2-a9c2-64d62c12dc42&acdnat=1528115323_108db79b95699aaa909141d0bdc961ec (Accessed: 4 June 2018).

Daggy, J. *et al.* (2010) 'Using no-show modeling to improve clinic performance', *Article Health Informatics Journal*, 16(4), pp. 246–259. doi: 10.1177/1460458210380521.

Devasahay, S. R., Karpagam, S. and Ma, N. L. (2017) 'Predicting appointment misses in hospitals using data analytics.', *mHealth*. AME Publications, 3, p. 12. doi:

10.21037/mhealth.2017.03.03.

Dove, H. G. and Schneider, K. C. (1981) 'The usefulness of patients' individual characteristics in predicting no-shows in outpatient clinics', *Medical care*, 19(7), pp. 734–40. doi: 10.1097/00005650-198107000-00004.

George, A. and Rubin, G. (2003) 'Non-attendance in general practice: a systematic review and its implications for access to primary health care', *Family Practice*, 20(2). doi: 10.1093/fampra/20.2.178.

Glowacka, K. J., Henry, R. M. and May, J. H. (2009) 'A Hybrid Data Mining/Simulation Approach for Modelling Outpatient No-Shows in Clinic Scheduling', *The Journal of the Operational Research Society*. Palgrave Macmillan JournalsOperational Research Society, pp. 1056–1068. doi: 10.2307/40206832.

Hosmer, D. W. *et al.* (1997) 'A comparison of goodness-of-fit tests for the logistic regression model.', *Statistics in medicine*, 16(9), pp. 965–80. Available at: http://www.ncbi.nlm.nih.gov/pubmed/9160492.

Hosmer, D. W. and Lemeshow, S. (2000) *Applied Logistic Regression*, *Wiley Series in Probability and Sattistics*. doi: 10.2307/2074954.

Hosmer DW, Lemeshow S, S. R. (2013) *Applied Logistic Regression, 3rd Edition*, *Wiley*. doi: 10.1002/9781118548387.

Huang, Y. and Hanauer, D. A. (2014) 'Patient no-show predictive model development using multiple data sources for an effective overbooking approach.', *Applied clinical informatics*. Schattauer Publishers, 5(3), pp. 836–60. doi: 10.4338/ACI-2014-04-RA-0026.

Huang, Y. and Zuniga, P. (2012a) 'Dynamic overbooking scheduling system to improve patient access', *Source: The Journal of the Operational Research Society Journal of the Operational Research Society Journal of the Operational Research Society*, 63(63), pp. 810–820. Available at: http://www.jstor.org/stable/41508617 (Accessed: 11 January 2018).

Huang, Y. and Zuniga, P. (2012b) 'Dynamic overbooking scheduling system to improve patient access', *Source: The Journal of the Operational Research Society Journal of the Operational*

Research Society Journal of the Operational Research Society, 63(63), pp. 810–820. Available at: http://www.jstor.org/stable/41508617.

Ivan, M.-L. and Velicanu, M. (2015) 'Healthcare Industry Improvement with Business Intelligence', *Informatica Economica*, 20(2/2015), pp. 81–89. doi: 10.12948/issn14531305/19.2.2015.08.

Junod Perron, N. *et al.* (2013) 'Text-messaging versus telephone reminders to reduce missed appointments in an academic primary care clinic: a randomized controlled trial.', *BMC health services research*. BioMed Central, 13, p. 125. doi: 10.1186/1472-6963-13-125.

Kramer, A. A. and Zimmerman, J. E. (2007) 'Assessing the calibration of mortality benchmarks in critical care: The Hosmer-Lemeshow test revisited*', *Critical Care Medicine*, 35(9), pp. 2052–2056. doi: 10.1097/01.CCM.0000275267.64078.B0.

LaGanga, L. R. and Lawrence, S. R. (2007) 'Clinic Overbooking to Improve Patient Access and Increase Provider Productivity', *Decision Sciences*. Wiley/Blackwell (10.1111), 38(2), pp. 251–276. doi: 10.1111/j.1540-5915.2007.00158.x.

LaGanga, L. R. and Lawrence, S. R. (2015) 'APPOINTMENT SCHEDULING WITH OVERBOOKING TO MITIGATE PRODUCTIVITY LOSS FROM NO-SHOWS'. Available at: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.530.8882&rep=rep1&type=pdf.

Law, A. M. (1991) *Simulation modeling and analysis*. Second Edi. Available at: https://ie.u-ryukyu.ac.jp/~asharif/pukiwiki/index.php?plugin=attach&pcmd=open&file=SIMULATION MODULING %26 ANALYSIS 1.pdf&refer=%A5%B7%A5□%E5%A5盞%A5%B7%A5%E7%A5%F3 (Accessed: 13 January 2018).

Luhn, H. P. (1958) 'A Business Intelligence System', *IBM Journal of Research and Development*, 2(4), pp. 314–319. doi: 10.1147/rd.24.0314.

Mbada, C. E. *et al.* (2013) 'Impact of missed appointments for out-patient physiotherapy on cost, efficiency, and patients' recovery', *Hong Kong Physiotherapy Journal*. No longer published by Elsevier, 31(1), pp. 30–35. doi: 10.1016/J.HKPJ.2012.12.001.

Moberly, T. (2014) ‘Text messaging and improved data help cut missed and cancelled appointments in London.’, *BMJ (Clinical research ed.)*. British Medical Journal Publishing Group, 348, p. g1840. doi: 10.1136/BMJ.G1840.

Neal, R. D. *et al.* (2005) ‘Reasons for and consequences of missed appointments in general practice in the UK: questionnaire survey and prospective review of medical records.’, *BMC family practice*. BioMed Central, 6, p. 47. doi: 10.1186/1471-2296-6-47.

Negash, S. (2004) ‘Communications of the Association for Information Systems Business Intelligence BUSINESS INTELLIGENCE’, *Communications of the Association for Information Systems*, 13(15), pp. 177–195. Available at: http://aisel.aisnet.org/cais (Accessed: 13 January 2018).

Parmar, V. *et al.* (2009) ‘The online outpatient booking system “Choose and Book” improves attendance rates at an audiology clinic: a comparative audit’, *Journal of Innovation in Health Informatics*, 17(3), pp. 183–186. doi: 10.14236/jhi.v17i3.733.

Sibte, S., Abidi, R. and Yusoff, Z. (1998) ‘Data-Driven Healthcare Management: From a Philosophy to an Info-Structure’. Available at: https://pdfs.semanticscholar.org/daae/af48a6a15c265e939ee4210eca0ca51dc012.pdf?_ga=2.217045142.1214599127.1528555047-1719046200.1528555047.

Siddiqui, Z. and Rashid, R. (2013) ‘Cancellations and patient access to physicians: ZocDoc and the evolution of e-medicine.’, *Dermatology online journal*, 19(4), p. 14. Available at: http://www.ncbi.nlm.nih.gov/pubmed/24021373.

Walters, B. A. and Danis, K. (2003) ‘Patient Online at Dartmouth-Hitchcock - interactive patient care web site.’, *AMIA ... Annual Symposium proceedings. AMIA Symposium*. American Medical Informatics Association, 2003, p. 1044. Available at: http://www.ncbi.nlm.nih.gov/pubmed/14728547.

Wang, L. (2005) *Support vector machines : theory and applications*. Springer. Available at: https://books.google.ie/books?hl=en&lr=&id=uTzMPJjVjsMC&oi=fnd&pg=PA1&dq=support+vector+machine+introduction&ots=GEBMer2Hg5&sig=HPawbq2VeKCZ0Cz4npB5TJaXOMc&redir_esc=y#v=onepage&q=support vector machine introduction&f=false.