Predictive Risk Modelling for Hospital Readmissions

A thesis

submitted in partial fulfilment

of the requirements for the degree

of

Master of Science

at

The University of Waikato

by

Claire Elizabeth Forsythe

THE UNIVERSITY OF
WAIKATO
Te Whare Wānanga o Waikato

University of Waikato

2014

Abstract

This thesis is concerned with developing a predictive risk model to identify patients that are at high risk of readmission to hospital. Such a model should have a desirable level of predictive accuracy but also, should be financially beneficial to the DHB. Logistic regression and Naive Bayes probabilistic classification methods were both considered to build the predictive model. Performance measures such as the positive predictive value (PPV) and cost savings analysis were used to find the optimal days between initial admission and readmission and the optimal threshold for prediction of high risk patients. This study is concerned with Waikato District Health Board (DHB) domiciled patients discharged between 1 July 2009 and 31 October 2013. The dataset includes information about the patients initial admission and the response variable is whether a readmission occurred or not.

Using logistic regression, this study found the model that fits the data best includes 21 variables that contain information about the patients initial admission. The two classification methods used produce a risk probability between 0 and 1 for each patient in the study. The logistic regression model performance was better than Naive Bayes as shown by the PPV (the proportion of patients correctly identified as at risk over the total at risk). The 56 day readmission data PPV at a risk threshold of 0.5 for the logistic regression was 48.4% and 30.8% for Naive Bayes. Analysis of the PPV identifies the risk threshold level of 0.5 and readmission period of 56 days as optimal predictive criteria in this study. Cost savings analysis also supports the 56 day model with an intervention cost of $500. The 0.5 cut off point in the 56 day model identifies a reasonable number of patients at risk for intervention, approximately 3,000, which equates to about 2 patients at risk per day over the period of this analysis.

This analysis found the optimal model for predicting patients at risk of readmission is a logistic regression model using 56 day readmission data and a risk threshold of 0.5. A key recommendation of this study is that the DHB needs to introduce a method that correctly flags patient admissions. The model

should be used on a trial basis at the DHB to see how accurate it performs.

Acknowledgements

I would like to express my sincere thanks to my supervisor Dr. Chaitanya Joshi, Department of Statistics, University of Waikato. The time and knowledge he has given me throughout this thesis is greatly appreciated. We spent a lot of time deliberating what to do and I know we were both very relieved to find such an interesting topic. This thesis evolved a lot over time and Chaitanya was available weekly to help me through it all.

I am very grateful to the Waikato District Health Board for allowing me to undertake this study in the DHB. Thanks to their incredible support and advice I was able to create this interesting model. Particular thanks to Paul Taumanu, Jan Adams, Paul Reeve and Linda Irving for their time, support and encouragement.

A massive thank you to my parents Dave and Sue for the support, encouragement and love they have always given me on many, many levels. Thanks to all of my family and friends for the cheerleading, advice and many forms of encouragement, you know who you are and I really appreciate it!

Finally, this thesis would not have been completed without the love, patience, and inspiration from Ken Tsang-Yum. Thanks for the encouragement and tolerance during this journey.

Tausisi I mea aupito aluga.

Contents

1 Introduction **1**

1.1 Aims of the Study 1

1.2 Research Contributions 2

1.3 Overview of Chapters 3

2 Logistic regression **5**

2.1 Notation and Coding for this Study 6

2.2 Exponential Family and Generalised Linear Models 7

2.2.1 Exponential Family 7

2.2.2 Binomial Distribution 8

2.2.3 Multinomial Distribution 8

2.2.4 Generalised Linear Models 9

2.3 Binary Variables and Logistic Regression 11

2.3.1 Probability Distributions 11

2.3.2 Generalised Linear Models 12

2.3.3 Logistic Model 12

2.3.4 General Logistic Regression Model 13

2.4 Parameter Estimation 14

2.4.1 Deviance 14

2.4.2 Akaike Information Criterion 15

2.4.3 Residual Deviance 16

2.5 Model Interpretation 17

2.5.1 Coefficients 17

2.5.2 p-values 17

2.5.3 Odds Ratios 17

2.6 Example of the Logistic Model in Predictive Modelling 17

3 Naive Bayes **19**

3.1 Probabilistic Reasoning 19

3.1.1 Conditional Probability and Bayes Theorem 19

3.1.2 Independence 20

3.1.3 Conditional Independence 21

3.2 Graphs . . . 21
3.3 Belief Networks . . . 21
3.3.1 Belief Networks . . . 22
3.4 Discrete Distributions . . . 23
3.5 Naive Bayes . . . 23
3.5.1 Conditional Independence . . . 23
3.5.2 Estimation Using the Maximum A Posteriori Probability 25
3.6 Logistic Regression versus Naive Bayes . . . 26

4 Readmissions and Predictive Risk Modelling 27
4.1 Readmissions . . . 27
4.1.1 Why Model Risk? . . . 28
4.1.2 28 Day Readmissions in New Zealand . . . 29
4.2 Readmission risk studies . . . 30
4.2.1 International studies . . . 30
4.2.2 Patients at Risk of Readmitting Within 30 Days (PARR-30) . . . 34
4.2.3 New Zealand Studies . . . 36

5 Data 38
5.1 Waikato District Health Board . . . 38
5.2 Data . . . 38
5.2.1 MoH Framework . . . 40
5.2.2 Variables . . . 41
5.3 Days Between Initial Admission and Readmission . . . 49
5.4 Actual Total Cost of Readmission . . . 50

6 Data Analysis 51
6.1 Modelling Data . . . 51
6.1.1 Logistic Regression . . . 51
6.1.2 Bayesian Belief Networks . . . 52
6.2 Model Performance Measures . . . 52
6.2.1 Positive Predictive Values . . . 52
6.2.2 Sensitivity . . . 53
6.2.3 Specificity . . . 53
6.2.4 Receiver Operating Characteristic Curve . . . 54
6.2.5 Summary of Performance Measures . . . 54
6.3 Cross Validation . . . 55
6.4 Cost Analysis . . . 56
6.5 Risk Band Table . . . 57
6.5.1 Risk Threshold . . . 57

6.5.2 Risk Band Table 58

7 Results **60**

7.1 Model Selection 61

7.1.1 Final model 64

7.2 Final model analysis 66

7.2.1 14 days 69

7.2.2 28 days 71

7.2.3 42 days 72

7.2.4 56 days 73

7.2.5 84, 182 and 365 days 74

7.2.6 Risk Threshold and Readmission Days Summary . . . 77

7.3 Naive Bayes 77

7.4 Cost Analysis 79

7.4.1 100% Readmission Reduction Cost Analysis 81

7.4.2 10%, 20% and 50% Readmission Reduction Cost Analysis 85

7.5 Conclusions 86

8 Discussion and conclusions **87**

8.1 Discussion 88

8.2 Recommendations 91

8.2.1 How to Measure a Readmission 91

8.2.2 Investigate the Efficacy of Intervention 92

8.2.3 Cost Analysis 92

8.2.4 Social Factors 93

8.3 Concluding Remarks 94

References 96

List of Figures

2.1 Example of the Logit Curve 13

3.1 Example of a Directed Acyclic Graph 24

7.1 Positive Predictive Value of all Logistic Regression models . . 70

7.2 False Negative Rate of all Logistic Regression models 70

7.3 Sensitivity of all Logistic Regression models 75

7.4 Specificity of all Logistic Regression models 75

7.5 Positive Predictive Value for Logistic Regression and Naive Bayes models 80

7.6 Sensitivity for Logistic Regression and Naive Bayes models . . 81

7.7 Average cost of a readmission for readmissions only 82

7.8 Cost savings for Logistic Regression models ($500 intervention) 83

7.10 Cost savings for Logistic Regression models ($1000 intervention) versus PPV 83

7.9 Cost savings for Logistic Regression models ($1000 intervention) 84

7.11 56 Day Readmission Model Cost Analysis for 10%, 20% and 50% Readmission Reductions 84

List of Tables

5.1 28 day readmission dataset readmission rates by Fiscal Year . 43

5.2 28 day readmission dataset readmission rates by Hospital . . . 44

5.3 28 day readmission dataset readmission rates by Specialty Cluster 44

5.4 28 day readmission dataset readmission rates by LOS 45

5.5 28 day readmission dataset readmission rates by Sex 45

5.6 28 day readmission dataset readmission rates by TLA 46

5.7 28 day readmission dataset readmission rates by Deprivation Score 47

5.8 28 day readmission dataset readmission rates by CCI weight group 48

7.1 Generalised Linear Model Selection 62

7.2 Logistic Regression Model risk band table 28 days 66

7.3 Logistic Regression Model total at risk and PPV by risk threshold and 14, 28, 42 and 56 readmission days 68

7.4 Logistic Regression Model total at risk and PPV by risk threshold for 84, 182 and 365 readmission days 69

7.5 Logistic Regression Model risk band table 56 days 73

7.6 Naive Bayes Model total at risk and PPV by risk threshold and 14, 28, 42 and 56 readmission days 79

Abbreviations

CCI: Charlson Comorbidity Index

DRG: Diagnostic Related Group

DHB: District Health Board

ED: Emergency Department

GLM: Generalised Linear Models

GP: General Practitioner

LOS: Length of Stay

MAP: Maximum A Posteriori

MoH: Ministry of Health

NHS: National Health Service

NMDS: National Minimum Dataset

NPV: Negative Predictive Value

OR: Odds Ratio

PARR-30: Patients at Risk of Readmitting within 30 days

PPV: Positive Predictive Value

ROC: Receiver Operating Curve

TLA: Territorial Local Authority

UK: United Kingdom

US: United States

Chapter 1

Introduction

A readmission is an acute, unplanned admission within a defined period of time of a previous admission.

Readmission rates are a well established health quality measure in New Zealand and internationally as Government health sectors and hospital managers regard it as a good performance measure. This is because some readmissions that do occur are avoidable and, if data is modelled correctly, groups of patients at high risk of readmission are identifiable. However what the data qualifies as a readmission in a large dataset may not actually be, for the individual person, an acute readmission related to the initial episode.

The Ministry of Health (MoH) reports use 28 days between the initial admission and the acute readmission as their optimal readmission time period. Many in the health sector regard this window as the most likely time frame that the two events are related. A longer time frame would increase the chances of picking up admissions unrelated to the initial admission.

1.1 Aims of the Study

This thesis is concerned with developing a predictive risk model to identify patients that are at high risk of readmission to hospital. A range of different

criteria is considered to find the optimal model for predicting patients at risk of readmission.

Statistical criteria such as coefficient p-values, odds ratios, model AIC and residual deviance and performance measures such as PPV and sensitivity are used to find the best model. Using these methods, the explanatory variables that fit the data best can be found. Predictive accuracy was measured using predictive criteria such as the optimal number of days between the initial discharge and readmission and risk threshold level are analysed. This criteria effects the number of patients that identified at risk, the total intervention costs and the potential savings from model utilisation. Cost analysis by using the actual cost of readmissions is used to calculate the possible savings if the model is implemented. The cost analysis and the PPV are the main methods used to find the best risk threshold and days between initial discharge and readmission.

In this thesis we set out to clarify what the ideal strategy is that saves the DHB the most money considering the criteria above. Although the MoH reports focus on the 28 day period; this thesis sets out to test whether that this is, in fact, the optimal time period and if not, what period should we focus on for predictive modelling purposes.

1.2 Research Contributions

In this thesis we set out to build a predictive risk model to identify patients at risk of readmission within a defined period of time. This unique model was developed for the Waikato DHB to identify patients for potential intervention. There are not many studies in New Zealand that have successfully implemented a model like this.

We tried to investigate the predictive criteria such as the number of readmission days and the risk threshold. We also looked into the success rate of intervention on high risk patients and used cost analysis to investigate potential savings to the DHB if the model were to be implemented.

1.3 Overview of Chapters

The thesis is organised as follows. In Chapter 2, the theory of logistic regression models is introduced. This chapter explains the general theory and methodology behind the models. It also explains the estimation techniques and how to interpret the logistic regression model output.

Chapter 3, similar to Chapter 2, explains the theory behind Naive Bayes, another classification model used in this thesis. The general ideas behind conditional probability, Bayes theorem, directed acyclic graphs and parameter estimation are explained in this chapter.

The background information into why readmissions and predictive risk modelling is important is found in Chapter 4. This chapter introduces the meaning of the term readmission and why it is important to analyse readmission risk at the DHB. It also summarises risk prediction studies that use different readmission periods, performance measures and explanatory variables to predict a patients risk of readmission.

In Chapter 5, the data that is used in this model from the Waikato DHB is explained. The possible explanatory variables are described as well as the data used for other analysis in this thesis.

The implementation of the models, performance measures and other tech-

niques used in this study are explained in Chapter 6. The classification models that are used are described as well as how this study will attempt to find the optimal model for risk prediction. It also explains how the models are validated, how the risk band tables are created and what cost analysis is performed.

The results of the predictive risk modelling is reported in Chapter 7. The reasons for including the explanatory variables used in the final logistic regression model are described. This chapter finds the optimal model in terms of risk threshold and readmission days. Logistic regression and Naive Bayes models are compared to see which one has better predictive power. Cost analysis results are analysed to confirm the optimal model.

The thesis is concluded in Chapter 8. The main findings of the study are summarised. The meaning of the proposed model is discussed and recommendations are made for future work.

Chapter 2

Logistic regression

There are two main types of classifiers, probabilistic or deterministic, that are used to classify observations into groups. Deterministic classifiers, which are not used in the this thesis, classify observations into the best fitting class (one group or another). Probabilistic classifiers are more informative than deterministic classifiers as the output of these models is the probability between 0 and 1 of an observation being a member of one class or another. An advantage of using these models, which is illustrated later in this thesis, is we can manipulate the threshold at which an observation is classed as one outcome or another. In this thesis we deal with two forms of probabilistic classifiers; Logistic Regression and Naive Bayes. Logistic regression is a special case of the generalised linear models and are discussed in this chapter. Logistic regression is a simple tool commonly used in health research for prediction as it forms linear combination of explanatory variables to predict an outcome (Perlis, 2013). Naive Bayes classification is discussed in the following chapter.

In this study our response variable is binary so we cannot use linear regression. The categorical outcome in our data is either a patient had a readmission or no readmission (0 or 1), which is not a continuous variable. This means this response cannot be modelled with a straight line so we need to use logistic regression to fit a model to our data.

In this chapter the basic theory regarding logistic regression used in this thesis is summarised. This theory is taken from the book by Dobson & Barnett (2008).

2.1 Notation and Coding for this Study

The observed values described in this thesis are denoted by lower case letters, $y_1, y_2, \ldots, y_n$ and are regarded as the realisation of the random variables which are denoted by upper case letters, $Y_1, Y_2, \ldots, Y_n$.

For the models described in this thesis the equation linking each response variable Y and a set of explanatory variables $x_1, x_2, \ldots, x_n$ has the form

$$g[E(Y)] = \beta_0 + \beta_1 x_1 + \ldots + \beta_m x_m.$$

The matrix notation for the response variables $Y_1, \ldots, Y_n$ is

$$g[E(Y)] = \boldsymbol{X}\boldsymbol{\beta}$$

where $\boldsymbol{y} = \begin{bmatrix} Y_1 \\ \vdots \\ Y_N \end{bmatrix}$ is a vector of responses,

$g[E(\boldsymbol{y})] = \begin{bmatrix} g[E(Y_1)] \\ \vdots \\ g[E(Y_N)] \end{bmatrix}$ denotes a vector of functions of the terms $E(Y_i)$ (with same g for every element),

$\boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_N \end{bmatrix}$ is a vector of parameters,

and $\boldsymbol{X}$ is a matrix whose elements are constants representing the different levels of categorical explanatory variables. For the categorical explanatory variables there are parameters for different levels of a factor. The corresponding elements of $\boldsymbol{X}$ are chosen to exclude or include the appropriate parameters

for each observation; they are called *dummy variables* or indicator variables (if there are only two categories 0 and 1).

If there are p parameters in a model and we have N observations then:

1. $\boldsymbol{y}$ is a $N \times 1$ random vector
2. $\boldsymbol{\beta}$ is a $p \times 1$ vector of parameters
3. $\boldsymbol{X}$ Is a $N \times p$ matrix of known constants. Also known as the design matrix.
4. And $\boldsymbol{X\beta}$ is the linear component of the model.

2.2 Exponential Family and Generalised Linear Models

In this thesis the response variable follows the binomial distribution and the relationship between the response and explanatory variables is not in simple linear form. The Exponential family of distributions are a class of probability distributions which share a common mathematical form. Many common distributions, such as the normal and binomial, belong to the exponential family. The theory behind Generalised linear models is based around the exponential family.

2.2.1 Exponential Family

For a single random variable Y whose probability distribution depends on a single parameter θ. The distribution belongs to the exponential family if it can be written in the form

$$f(y;\theta) = s(y)t(\theta)e^{(a(y)b(\theta)}, \tag{2.1}$$

where a, b, s and t are known functions. The symmetry in y and θ is emphasised in the following equation

$$f(y;\theta) = \exp\left[a(y)b(\theta) + c(\theta) + d(y)\right], \tag{2.2}$$

where $s(y) = \exp d(y)$ and $t(\theta) = \exp c(\theta)$.

2.2.2 Binomial Distribution

This thesis deals with a series of binary events. The observations in the dataset each have a possibility of two outcomes: readmission or no readmission. Let the random variable Y be the number of readmissions in n independent trials in which the probability of a readmission, π, is equal for all trials. Then Y has a Binomial distribution with the probability density function

$$f(y;\pi) = \binom{n}{y}\pi^y(1-\pi)^{n-y}, \tag{2.3}$$

where y takes the values $0, 1, 2, ..., n$ and

$$\binom{n}{y} = \frac{n!}{y!(n-y)!}. \tag{2.4}$$

which is the binomial coefficient denoted by $Y \sim \text{Bin}(n, \pi)$. The parameter of interest is π and we assume that n is known. The probability function can be written in the canonical form according to equation (2.2)

$$f(y;\pi) = \exp\left[y\log\pi - y\log(1-\pi) + n\log(1-\pi) + \log\binom{n}{y}\right], \tag{2.5}$$

with $b(\pi) = \log\pi - \log(1-\pi) = \log\frac{\pi}{1-\pi}$.

Note that when $n = 1$ the Bernoulli distribution is a special case of the binomial distribution. The Bernoulli distribution is simply Binomial$(1, p)$.

2.2.3 Multinomial Distribution

The response variable in the dataset in this study in binomial and so are some of the explanatory variables. However many of the explanatory variables have

more than two categories. The multinomial distribution can be used to compute the probabilities in situations in which there are more than two possible outcomes.

Consider random variable Y with J categories. Let $\pi_1, ...\pi_J$ denote the respective probabilities, with $\pi_1 + \pi_2 + .. + \pi_J = 1$. If there are n independent observations of Y which result in y_1 outcomes in category 1, y_2 outcomes in category 2 and so on then $\sum_{j=1}^{J} y_j = n$.
The multinomial distribution, $M(n, \pi_1, ..., \pi_j)$, is

$$f(y|n) = \frac{n!}{y_1!y_2!, ..., y_J!}\pi_1^{y_1}\pi_2^{y_2}...\pi_J^{y_J} \tag{2.6}$$

Note that Equation (2.6) is not a member of the exponential family of distributions but its relationship with the Poisson distribution ensures that generalised linear modelling is appropriate. The Multinomial distribution can be regarded as the joint distribution of Poisson random variables, conditional upon their sum n. For further details on the relationship see the Chapter 8 of Dobson & Barnett (2008).

2.2.4 Generalised Linear Models

The generalised linear model is defined as the set of independent random variables $Y_1, \ldots, Y_N$ each with a distribution from the exponential family and has the following properties:

1. The distribution of each Y_i has the canonical form and depends on a single parameter θ_i (the θ_is do not all have to be the same); thus:

$$f(y_i; \theta_i) = \exp[y_i b_i(\theta_i) + c_i(\theta_i) + d_i(y_i)].$$

   The Y_i's in this thesis are in the form of equation (2.5).

2. The distributions of all Y_i's are of the same form, in our case all are Binomial, so that the subscripts on b, c and d are not needed.

Thus, the joint probability density function of $Y_1, \ldots, Y_N$ is

$$f(y_1, \ldots, y_N; \theta_1, \ldots, \theta_N) = \prod_{i=1}^{N} \exp[y_i b_i(\theta_i) + c_i(\theta_i) + d(y_i)] \tag{2.7}$$

$$= \exp[\sum_{i=1}^{N} y_i b_i(\theta_i) + \sum_{i=1}^{N} c_i(\theta_i) + \sum_{i=1}^{N} d(y_i)] \tag{2.8}$$

For the generalised linear model we use a transformation of μ_i such that

$$g(\mu_i) = \boldsymbol{x}_i^T \boldsymbol{\beta}$$

Where in the equation above:

- g is the *link function* which is a monotone differential function that is flat so it cannot increase for some values of μ_i and decrease for others.
- The vector $\boldsymbol{x}_i$ is a $p \times 1$ vector of the explanatory variables in the form

$$\boldsymbol{x_i} = \begin{bmatrix} x_{i1} \\ \vdots \\ x_{ip} \end{bmatrix} \quad \text{so} \quad \boldsymbol{x}_i^T = \begin{bmatrix} x_{i1} & \ldots & x_{ip} \end{bmatrix}.$$

  The vector $\boldsymbol{x_i}$ is the ith row in the design matrix $\boldsymbol{X}$.

- Lastly the vector of parameters, $\boldsymbol{\beta}$ is a $p \times 1$ vector $\boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}$.

So the generalised linear models used in this thesis have the following elements:

1. Response variables $Y_1, \ldots, Y_N$ which are assumed to share the same distribution from the exponential family;
2. A set of parameters β and explanatory variables

$$\boldsymbol{X} = \begin{bmatrix} \boldsymbol{x}_1^T \\ \vdots \\ \boldsymbol{x}_N^T \end{bmatrix} = \begin{bmatrix} x_{11} & \ldots & x_{1p} \\ \vdots & & \vdots \\ x_{N1} & & x_{Np} \end{bmatrix};$$

3. Lastly a monotone link function g such that

$$g(\mu_i) = \boldsymbol{x}_i^T \boldsymbol{\beta}$$

where $\mu_i = E(Y_i)$.

2.3 Binary Variables and Logistic Regression

2.3.1 Probability Distributions

In this thesis the response variable is a *binary random variable* which is defined as

$$Z = \begin{cases} 1 & \text{if the outcome is a readmission;} \\ 0 & \text{if the outcome is not a readmission.} \end{cases}$$

with the probabilities $\Pr(Z = 1) = \pi$ and $\Pr(Z = 0) = 1 - \pi$ which follow the Bernoulli distribution $\mathrm{B}(\pi)$. If there are n such random variables $Z_1, ..., Z_n$, which are independent with $\Pr(Z_j = 1) = \pi_j$, then their joint probability is in the canonical form of the exponential family of distributions (equation 2.2):

$$\prod_{j=1}^{n} \pi_j^{z_j}(1-\pi_j)^{1-z_j} = \exp\left[\sum_{j=1}^{n} z_j \log\left(\frac{\pi_j}{1-\pi_j}\right) + \sum_{j=1}^{n} \log(1-\pi_j)\right], \quad (2.9)$$

When the π_j's are all equal, we can define

$$Y = \sum_{j=1}^{n} Z_j$$

so that Y is the number of readmissions in n observations. The random variable Y has the distribution $\mathrm{Bin}(n, \pi)$ with a Binomial probability density function as seen in equation (2.3).

If $Y \sim \mathrm{Bin}(n_i, \pi_i)$, the log-likelihood function is:

$$l(\pi_1, ..., \pi_N, y_1, ..., y_N) = \sum_{i=1}^{N} \left[y_i \log\left(\frac{\pi_i}{1-\pi_i}\right) + n_i \log(1-\pi_i) + \log\binom{n_i}{y_i} \right]. \quad (2.10)$$

2.3.2 Generalised Linear Models

Want to describe the proportion of readmissions, $P_i = Y_i/n_i$, in each subgroup in terms of factor levels and other explanatory variables which characterize the subgroup. As $E(Y_i) = n_i\pi_i$ and so $E(P_i) = \pi_i$ we model the probabilities π_i using the equation

$$g(\pi_i) = \boldsymbol{x}_i^T\boldsymbol{\beta},$$

where $\boldsymbol{x}_i$ is a vector of explanatory variables (dummy variables for factor levels and measured values for covariates), $\boldsymbol{\beta}$ is a vector of parameters and g is a link function.

To ensure π is restricted to between 0 and 1 it is often modeled using a cumulative probability distribution

$$\pi = \int_{-\infty}^{t} f(s)ds,$$

where $f(s) \geq 0$ and $\int_{-\infty}^{\infty} f(s)ds = 1$. The probability density function $f(s)$ is called the *tolerance distribution*.

2.3.3 Logistic Model

The *logistic model*, also known as *logit model* is a special case of the generalised linear model and is known to be reasonably easy to compute. It has the probability density function, the *tolerance distribution*:

$$f(s) = \frac{\beta_2 \exp(\beta_1 + \beta_2 s)}{[1 + \exp(\beta_1 + \beta_2 s)]^2}.$$

To ensure that the probability, π lies between 0 and 1 we model it using the cumulative probability distribution

$$\pi = \int_{-\infty}^{x} f(s)ds = \frac{\exp(\beta_1 + \beta_2 x)}{1 + \exp(\beta_1 + \beta_2 x)}$$

This results in the link function

$$\log\left(\frac{\pi}{1-\pi}\right) = \beta_1 + \beta_2 x. \tag{2.11}$$

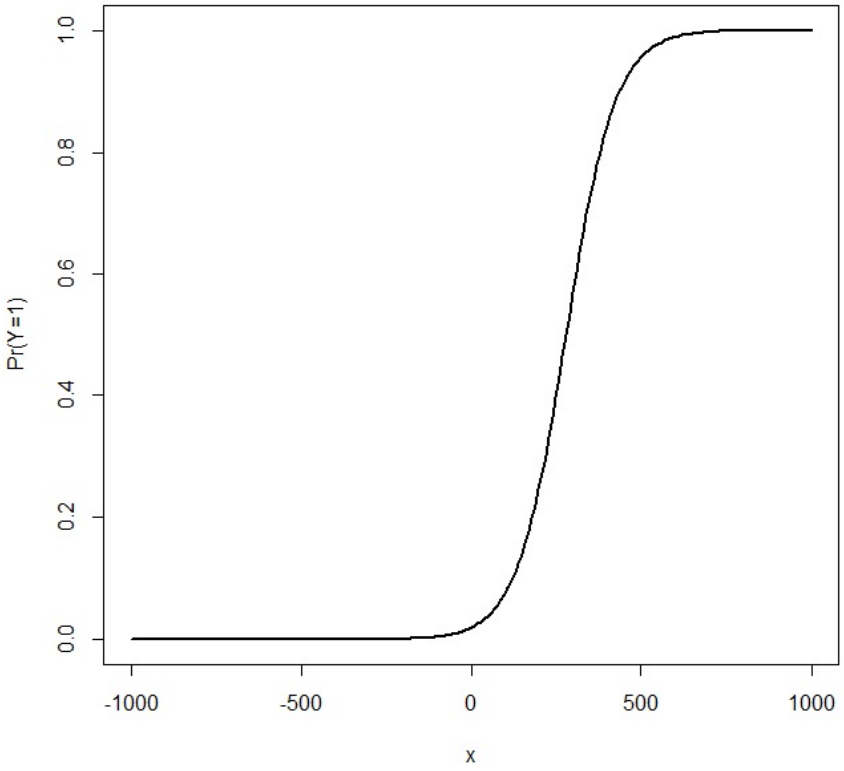

Figure 2.1: Example of the Logit Curve

The term $\log[\pi/(1-\pi]$ is sometimes called the *logit function* and it has natural interpretation as the logarithm of the odds (which is used to calculate the odds ratios for the coefficients used later in this thesis). The logistic model is used to model the binary data in this thesis. The logit curve follows a sigmoid "S" shape as shown in Figure (2.1).

By fitting the logistic model to our binomial data we get the log-likelihood fucntion based on equation (2.10) is

$$l = \sum_{i=1}^{N} \left[y_i \left(\beta_1 + \beta_2 x_i\right) + n_i \log(1 + \exp(\beta_1 + \beta_2 x_i)) + \log \binom{n_i}{y_i} \right].$$

2.3.4 General Logistic Regression Model

The general logistic regression model

$$\text{logit}\pi_i = \log\left(\frac{\pi_i}{1-\pi_i}\right) = \boldsymbol{x}_i^T \boldsymbol{\beta},$$

where $\boldsymbol{x}_i$ is vector of continuous measurements corresponding to covariates and dummy variables corresponding to factor levels and $\boldsymbol{\beta}$ is the parameter vector. This model is very widely used for analysing data involving binary or Binomial responses and several explanatory variables. It provides a powerful technique analogous to multiple regression and ANOVA for continuous responses.

2.4 Parameter Estimation

For generalised linear models the maximum likelihood estimates are obtained by solving the iterative weighted least squares procedure until the log-likelihood converges to a maximum. This i depicted in the following equation

$$\Im^{(m-1)}\mathbf{b}^{(m)} = \Im^{(m-1)}\mathbf{b}^{(m-1)} + \mathbf{U}^{(m-1)} \tag{2.12}$$

where $\Im$ is the information matrix, $\mathbf{U}$ is the score function which is the derivative of the log likelihood function and $\mathbf{b}^{(m)}$ is the vector of estimates of the parameters $\beta_1, \ldots, \beta_p$ at the mth iteration. The previous equation can be written as

$$\mathbf{X}^T\mathbf{W}\mathbf{X}\mathbf{b}^{(m)} = \mathbf{X}^T\mathbf{W}\mathbf{z}. \tag{2.13}$$

The method used to fit the generalised linear model to the data in this thesis is based on equation (2.13). Starting by using an initial approximation for $\mathbf{b}^{(0)}$ to evaluate $\mathbf{z}$ and $\mathbf{W}$, then equation (2.13) is solved to give $\mathbf{b}^{(1)}$, which then gives us better approximations for $\mathbf{z}$ and $\mathbf{W}$ until they converge. When the difference between $\mathbf{b}^{(m-1)}$ and $\mathbf{b}^{(m)}$ is small enough $\mathbf{b}^{(m)}$ is taken as the maximum likelihood estimate.

More details around how these equations are derived see the book by Dobson & Barnett (2008).

2.4.1 Deviance

The *Deviance* is a quality of fit statistic used to compare two models. It measures how the a model fits the data. A large deviance may indicate a poor fit

to the data. Typically as you add more variables to a model the deviance will decrease which indicates an improvement in fit.

It is calculated using the equation

$$D = 2\left[l(\mathbf{b}_{max}; y) - l(\mathbf{b}; \mathbf{y})\right]$$

where $l(\mathbf{b}_{max}; y)$ is the likelihood of the saturated model (this is the model with the maximum number of parameters that can be estimated used to compare to the model of interest) and $l(\mathbf{b}; \mathbf{y})$ is the maximum likelihood of the model of interest.

If the response variables are independent and $Y_i \sim Bin(n_i, \pi_i)$ then the log-likelihood function is

$$l(\beta; \mathbf{y}) = \sum_{i=1}^{N} \left[y_i \log \pi_i - y_i \log(1 - \pi_i) + n_i \log(1 - \pi_i) + \log \binom{n_i}{y_i} \right]$$

The deviance for the Binomial model is

$$D = 2 \sum_{i=1}^{N} \left[y_i \log \left(\frac{y_i}{\hat{y}_i} \right) + (n_i - y_i) \log \left(\frac{n_i - y_i}{n_i - \hat{y}_i} \right) \right] \tag{2.14}$$

Notice that D does not involve any nuisance parameters (like σ^2 for Normal response data), so goodness of fit can be assessed and hypotheses can be tested directly using the approximation

$$D \sim \chi^2(N - p),$$

where p is the number of parameters estimated and N the number of covariate patterns.

2.4.2 Akaike Information Criterion

Another goodness of fit statistic used in this thesis is the *Akaike information criterion* (AIC). This commonly used measure summarises how well a model fits the data as it measures the information that is lost when fitting a model

to the data. AIC is based on the log-likelihood function from the model, $l(\hat{\boldsymbol{\pi}}; \boldsymbol{y})$, with an adjustment for the number of parameters estimated and for the amount of observations:

$$AIC = -2l(\hat{\boldsymbol{\pi}}; \boldsymbol{y}) + 2p. \tag{2.15}$$

AIC discourages overfitting of models as it includes a "penalty" for the number of parameters included in the model. When fitting GLMs a smaller AIC value is preferable.

2.4.3 Residual Deviance

Deviance residuals are another goodness of fit measure for GLMs which correspond to the deviance, D. They are good for testing the adequacy of a model

$$d_k = \text{sign}(y_k - n_k\hat{\pi}_k)\left(2\left[y_k \log\left(\frac{y_k}{n_k\hat{\pi}_k}\right) + (n_k - y_k)\log\left(\frac{n_k - y_k}{n_k - n_k\hat{\pi}_k}\right)\right]\right)^{1/2} \tag{2.16}$$

where Y_k denotes the number of successes, n_k is the number of trials and $\hat{\pi}_k$ the estimated probability of success for the kth covariate pattern.

From equation (2.14), $\sum_{k=1}^{m} d_k^2 = D$, the deviance. Also the standardised deviance residuals are defined by

$$r_{Dk} = \frac{d_k}{\sqrt{1 - h_k}}$$

where h_k is the leverage obtained from the hat matrix.

Since the response variable used in this thesis is binary there are few distinct values of the residuals. This means normal probability plots may be relatively uninformative. For that reason we use the goodness of fit statistics the AIC and Residual Deviance.

2.5 Model Interpretation

2.5.1 Coefficients

The coefficients for each of the explanatory variables are estimated by the parameter estimation method described above using the equation (2.11). The coefficients are used to find the odds ratios as described below.

2.5.2 p-values

The p-values are found for each coefficient when performing logistic regression on the data. They are the probability of observing more extreme data (if the random process of the data were repeated) given that the null hypothesis is correct (that the coefficient is different from 0 or not).

2.5.3 Odds Ratios

The logit function used in the logistic regression in this study has the natural interpretation as the log of the odds, see equation (2.11). To get the value π which is the probability of "success", or in our case readmission, we take the exponential of the right hand side of the equation. This is because it is easier to interpret the explanatory variable effects in terms of odds ratios rather than parameters for β. So the odds is the ratio of the probability of the outcome, readmission, occurring over the odds of a readmission not occurring.

$$\frac{\pi}{1-\pi} = e^{\beta_1+\beta_2 x}. \tag{2.17}$$

2.6 Example of the Logistic Model in Predictive Modelling

The following is an example of how a logistic regression model can be used in predictive risk modelling. A logistic model for predicting patients at risk of readmission has a response variable for readmission, 0 or 1. Suppose we have

only four explanatory variables used in the model are age, sex, ethnicity and year of admission.

The coefficients are calculated using

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 \text{Age} + \beta_2 \text{Sex} + \beta_3 \text{Ethnicity} + \beta_4 \text{Year}$$

where π_i is the probability of a patient readmitting and

$$\frac{\pi_i}{1-\pi_i} = e^{\beta_0+\beta_1 \text{Age}+\beta_2 \text{Sex}+\beta_3 \text{Ethnicity}+\beta_4 \text{Year}}$$

is the odds ratio.

Chapter 3

Naive Bayes

The probabilistic models, Naive Bayes, are a simple form of Bayesian Belief Networks. The observations are classified as a binary variable using the probability generated by the logarithm for each observation. This chapter describes the basic theory of naive Bayes methods and is mainly based on the book Bayesian reasoning and machine learning Barber (2012).

Logistic regression does not model potential causal relationships between variables as all risk factors are treated as directly related to readmission risk (Nguefack-Tsague, 2011).The Naive Bayes classifier uses Bayes Theorem for probabilistic classification similar to logistic regression. It is assumed that the predictors are conditionally independent (Perlis, 2013).

3.1 Probabilistic Reasoning

3.1.1 Conditional Probability and Bayes Theorem

The probability of the event x conditioned on knowing event y (or the probability of x given y) is defined as

$$p(x \mid y) \equiv \frac{p(x, y)}{p(y)} \tag{3.1}$$

But if $p(y) = 0$ then $p(x \mid y)$ is not defined. We also know that $p(x, y) = p(y, x)$ therefore Bayes rule is defined as

$$p(x \mid y) = \frac{p(y \mid x)p(x)}{p(y)}. \tag{3.2}$$

In the context of statistical inference we typically have variable, θ, given that we have observed the data, D. We can rewrite Equation 3.2 as

$$p(\theta|D) = \frac{p(D|\theta)p(\theta)}{p(D)} = \frac{p(D|\theta)p(\theta)}{\int_\theta p(D|\theta)p(\theta)} \tag{3.3}$$

which shows we can infer the *posterior* distribution $p(\theta|D)$. Bayesian inference uses this equation for parameter estimation.

The *Most Probable A Posteriori* (MAP) setting is that which maximises the posterior. For a flat prior $p(\theta)$ is constant so the MAP solution is equivalent to the maximum likelihood.

3.1.2 Independence

Variables x and y are independent if knowing the state of one variable gives no extra information about the other variable. This can be written mathematically as

$$p(x, y) = p(x)p(y) \tag{3.4}$$

Provided that $p(x) \neq 0$ and $p(y) \neq 0$ independence of x and y is equivalent to

$$p(x|y) = p(x) \Leftrightarrow p(y|x) = p(y). \tag{3.5}$$

If $p(x|y) = p(x)$ for all states of x and y then the variables x and y are independent. If

$$p(x, y) = kf(x)g(y) \tag{3.6}$$

for some constant k, and positive functions $f(.)$ and $g(.)$ then x and y are independent and we write $x \perp\!\!\!\perp y$.

3.1.3 Conditional Independence

The following equation denotes that the two sets of variables X and Y are independent of each other provided we know the state of the set of variables Z.

$$X \perp\!\!\!\perp Y|Z \tag{3.7}$$

For conditional independence X and Y must be independent given *all* states of Z. This is depicted in the following equation

$$p(X \perp\!\!\!\perp Y|Z) = p(X|Z)p(Y|Z) \tag{3.8}$$

for all states of X, Y and Z.

3.2 Graphs

A graph consists of nodes and edges. Nodes are the variables included in the model and the edges are the links or the relationships between the nodes. A directed graph has edges that are directed, meaning they have an arrow in a single direction. An undirected graph has all edges undirected. Two variables will be independent if they are not linked by a path on the graph.

A *path* is the sequence of nodes that connects node A to node B in $A \mapsto B$.

A *acyclic* graph has directed path that does not return to the same node.

Directed Acyclic Graph (DAG) is a graph with directed edges between the nodes such that by following the path of nodes from one node to another along the direction of each edge no path will revisit a node.

In a DAG the *parents* are the nodes that lead to a specific node and the *children* are the nodes that follow from the specific node.

3.3 Belief Networks

Belief networks, also known as Bayesian belief networks, depict the independence assumptions made in a distribution. They are a way of applying Bayes

theorem to machine learning techniques and are a network of random variables and their conditional probabilities shown through directed links between variables.Belief networks are used because we do not want to make a large probability table with all the conditional probabilities between the variables which is very computationally intensive in large datasets. This is because computing the marginal probability requires summing 2^{N-1} states of other variables. So the marginal probability can be computed quickly we want to identify what variables are independent to each other.

3.3.1 Belief Networks

A belief network represents the factorisation of a distribution into conditional probabilities of variables dependent on parental variables.
A **belief network** is a distribution of the form:

$$p(x_1, ..., x_D) = \prod_{i=1}^{D} p(x_i \mid pa(x_i)) \tag{3.9}$$

where $pa(x_i)$ represent the *parental* variables of variable x_i. A belief network is a DAG with arrows pointing to the child nodes from the parental nodes.

The following properties describe some of the effects of conditioning or marginalising a variable in a belief network:

1. Imagine if A and B are the parents of C. $A \rightarrow C \leftarrow B$ then

$$p(A, B, C) = p(C|A, B)p(A)p(B) \tag{3.10}$$

   A and B are priori independent, both determining the effect of C.

2. Also in $A \rightarrow C \leftarrow B$ if you marganilise over C then it makes A and B independent.

3. For $A \rightarrow C \leftarrow B$ the conditioning on C makes A and B (graphically) dependent so that $p(A, B|C) \neq p(A|C)p(B|C)$

4. When $A \leftarrow C \rightarrow B$

   $p(A, B, C) = p(A|C)p(B|C)p(C)$

   In this graph C is the cause and A and B are the effects.

5. In $A \leftarrow C \rightarrow B$ if you marginalise over$C$ then it makes A and B dependent so that $p(A, B) \neq p(A)p(B)$

6. But for $A \leftarrow C \rightarrow B$ conditioning on C makes A and B independent. $p(A, B|C) = p(A|C)p(B|C)$

7. The following graphs all represent the same conditional independence structure $A \leftarrow C \rightarrow B$, $A \leftarrow C \leftarrow B$ and $A \rightarrow C \rightarrow B$.

3.4 Discrete Distributions

The response variable in this study is the *Binomial distribution*. It has two outcomes, readmission or no readmission. The probability of a success in n Bernoulli trials there will be k success states or readmissions.

$$p(y = k|\theta) = \binom{n}{k}\theta^k(1-\theta)^{n-k} \tag{3.11}$$

$$mean(y) = n\theta \qquad var(y) = n\theta(1-\theta)$$

Beta distribution is the conjugate prior for θ.

3.5 Naive Bayes

Naive Bayes is a special case of Bayesian Belief Networks. It is a simple method used to classify data.

3.5.1 Conditional Independence

For Naive Bayes we only use the intuitive concept of classification, being that we give a discrete label to an observation.

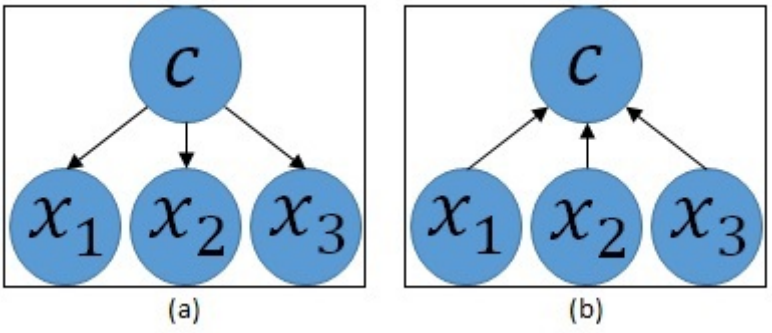

Figure 3.1: Example of a Directed Acyclic Graph

$$p(\boldsymbol{x}, c) = p(c) \prod_{i=1}^{D} p(x_i|c) \tag{3.12}$$

whose belief network is seen in Figure 3.1 (a). Coupled with a suitable choice for each conditional distribution $p(x_i|c)$ we then use Bayes rule to form a classifier for a novel input vector $\boldsymbol{x}^*$:

$$p(c|\boldsymbol{x}^*) = \frac{p(\boldsymbol{x}^*|c)p(c)}{p(\boldsymbol{x}^*)} = \frac{p(\boldsymbol{x}^*|c)p(c)}{\sum_c p(\boldsymbol{x}^*|c)p(c)} \tag{3.13}$$

In this thesis we only consider two classes $\text{dom}(c) = \{0, 1\}$.

Most of the theory regarding Naive Bayes describes c as the causes of x_i, as displayed in Figure 3.1 (a). In this study the explanatory variables x_i are the cause of c, the readmission, which can be seen in in Figure 3.1 (b). This means the conditional independence statement can be written as either $p(c|\boldsymbol{x}^*)$ or $p(\boldsymbol{x}^*|c)$. In this study there is an assumption that there is independence between the explanatory variables. The prior probability of the response variable, the readmissions, could be determined using the readmission rates from previous studies. The conditional probability for each of the explanatory variables is known. This study uses the maximum a posteriori (MAP) probability estimate to find the probability associated with each observation so we can attribute each to one of two classes, 0 or 1 (no readmission or readmission) which is described below.

3.5.2 Estimation Using the Maximum A Posteriori Probability

There are two ways to estimate the probability $p(\boldsymbol{x}^*|c)p(c)$ which is either Maximum Likelihood Estimation or Bayesian estimation. In this thesis we use Bayesian estimation which is described below.

For each predictor, x_i, and for every outcome c_k of the response, $X_i|c_k$ follows a Multinomial Distribution. Let $D_{ik} = x_i|c_k$, denote the data where each follows a Multinomial distribution

$$D_{ik} \sim \text{Multinomial}(N, \theta_{i1k}, \theta_{i2k}, \ldots, \theta_{iqk})$$

where q is the number of different values the variable x_i can take.

The likelihood of D is

$$f(D|\theta_{i1k}, \theta_{i2k}, \ldots, \theta_{iqk}) = \frac{N!}{x_{i1k}! \ldots x_{iqk}!} \prod_{j=1}^{q} (\theta_{ijk})^{x_{ijk}}$$

where $N = \sum_{j=1}^{q} x_{ijk}$.

The conjugate prior $(\theta_{i1k}, \ldots, \theta_{iqk}) \sim \text{Dirichlet}(q, \gamma_1, \ldots, \gamma_q)$

$$f(\theta_{i1k}, \ldots, \theta_{iqk}) = \frac{1}{B(\gamma_1, \ldots, \gamma_q)} \prod_{j=1}^{q} (\theta_{ijk})^{\gamma_j}$$

where $\theta_{ijk} > 0 \ \forall_j$ and $\sum_{j=1}^{q-1} \theta_{ijk} < 1$ and $\theta_{ijk} = 1 - \sum_{j=1}^{q-1} \theta_{ijk}$ and $B(\gamma_1, \ldots, \gamma_q)$ is the normalising constant.

The posterior follows a Dirichlet distribution

$$P(\theta_{i1k}, \ldots, \theta_{iqk}|x_i, c_k) = \text{Dirichlet}(q, x_{i1k} + \gamma_1, \ldots, x_{iqk} + \gamma_q)$$

So that the Maximum A Posteriori (MAP) is

$$\hat{\theta}_{ijk} = \underset{\theta}{\text{argmax}} \, P(\theta_{i1k}, \ldots, \theta_{iqk}|D).$$

The class variable (readmission or no readmission) follows a binomial distribution $c \sim \text{Binomial}(n, \pi)$ and the likelihood is

$$f(c_1|\pi) = \binom{n}{c_1} \pi^{c_1}(1-\pi)^{c_2}$$

where $c_1 + c_2 = N$. The conjugate prior $\pi \sim \text{Beta}(\alpha, \beta)$ is

$$f(\pi) = \frac{\pi^{\alpha-1}(1-\pi)^{\beta-1}}{B(\alpha, \beta)}$$

where $B(\alpha, \beta)$ is the normalising constant.

The posterior $P(\pi|c) = \text{Beta}(c_1 + \alpha, c_2 + \beta)$ so

$$\hat{\pi} = \underset{\pi}{\text{argmax}}\, P(\pi_1, \pi_2|c).$$

3.6 Logistic Regression versus Naive Bayes

Logistic regression is a discriminative classifier as it directly estimates the parameters of the distribution $P(c|x)$. In this case we can only model the response variable which is conditional on the predictor variables. Alternatively Naive Bayes is a generative classifier as it estimates parameters for $P(c)$ and $P(x_i|c)$. Using this classification technique we can simulate anything we like, the response or the predictor variables and so on. One method is not superior to the other so a purpose of this study is to compare the two models and decide which algorithm is the best classifier for predictive risk modelling.

Chapter 4

Readmissions and Predictive Risk Modelling

4.1 Readmissions

A *readmission* is "the next subsequent admission of a patient as an acute (that is, emergency or unplanned) admission within a defined period of time" (Rumball-Smith, 2009). Clinicians believe key factors for readmissions are a patients condition, level of frailty, age and level of social support.

Many argue the suitability of readmission rates as a quality measure in hospitals (Drozda, 2013). Many medical staff at Waikato District Health Board (DHB) question whether readmission rates are a good quality indicator as they expect patients to come back to hospital at some point which cannot, unfortunately, be stopped. But government health sectors and hospital managers internationally regard it as a good performance measure as some readmissions that do occur are avoidable and if data is modelled correctly groups of patients at high risk of readmission are identifiable (Rumball-Smith, 2009).

Unfortunately, a downfall of this measure is whether the readmission the data identifies, using common readmission query criteria, is actually an acute

readmission on the initial episode of care. What the data qualifies as a readmission in a large dataset may not actually be, for the individual person, an acute readmission related to the initial episode. For example a patient may have an elective surgery for a hip replacement and then readmit to hospital due to an emergency unrelated to that surgery such as a motor vehicle accident (MVA). The data would qualify that MVA as a readmission. Unfortunately, at Waikato DHB, information regarding the association between initial episodes and readmissions is not collected.

4.1.1 Why Model Risk?

Acute readmissions in New Zealand are a Ministry of Health (MoH) quarterly performance measure, also known as Ownership 8 (OS8): Acute Readmissions to Hospital. The MoH compare rates between DHBs four times per year and hospital management use these rates to gauge how they are doing in comparison to similar DHBs in New Zealand. They can also identify where problem areas may be through further data analysis.

According to the MoH "hospital unplanned acute readmission rates are a well-established measure of quality of care, efficiency, and appropriateness of discharge for hospital patients, particularly as a counter-measure to reduce a hospitals average length of stay. International experience is that shorter lengths of stay are correlated with higher rates of acute readmissions. Unplanned acute readmissions may imply a possible failure in patient management such as discharge too early, or inadequate support at home" (New Zealand Ministry of Health, 2012).

In the United Kingdom (UK) the National Health Service (NHS) actually proposes that hospitals should not be paid for acute readmissions within 30 days of a planned elective surgery (Billings, Blunt, Steventon, Georghiou,

Lewis & Bardsley, 2012). Readmission rates are an important monitoring tool for health outcomes in the UK. Similarly, in Australia the rates of unplanned readmissions within 28 days for specific surgical procedures are used as a performance indicator for comparison between hospitals and states (Australian Institute of Health and Welfare, 2014).

In the United States (US) there are serious implications on hospitals for readmissions as financial penalties are enforced meaning some hospitals do not receive government funding for selected readmissions (Drozda, 2013). A lot of research has gone into finding ways to reduce hospital readmissions resulting in longer patient length of stay as hospitals keep patients in hospital for longer because of the financial penalties they may endure in the US.

4.1.2 28 Day Readmissions in New Zealand

The MoH reports use 28 days between the initial admission and the acute readmission as their optimal readmission time. This idea is supported by many in the health sector as this window is regarded as the most likely time frame that the two events are related. The 28 day criteria maybe based on a compromise between false negatives and false positives. This is because if the time period is too short there will be an increase in false negatives (when the model predicts patients as low risk but they do readmit) and if the time period is too long then the false positive rate is likely to be higher (the rate of patients who are identified as high risk but do not readmit).

Rumball-Smith (2009) found that 43% of the studies they researched had a period of approximately one month between initial admission and acute readmission. This period is widely used in government bodies in countries such as Canada, Australia, New Zealand and the UK. A longer time frame would increase the chances of picking up admissions unrelated to the initial admission.

Robinson (2012) argue it may be dependent on the question that is being asked about the readmission. For example if one is more concerned with deficiencies in care then a shorter time frame maybe appropriate but if you were concerned with supporting the transition of patients into the community after a hospital stay then a longer time frame such as 12 months might be more useful.

There is potential to investigate if the 28 day time period is indeed the optimal time period between initial discharge and readmission so we can build th best model for predicting patients at high risk of readmitting. Different readmission periods are investigated in this thesis to find the optimal time between initial admission and readmission for predictive risk modelling.

4.2 Readmission risk studies

Predictive risk modelling is a case finding algorithm that attempts to identify patients at risk of a readmission. Research indicates these models have been used for some time and have developed significantly over time.

4.2.1 International studies

The National Health Service (NHS) in the United Kingdom (UK) have, for sometime, been developing predictive risk models. Case finding mechanisms are used to identify high risk patients to enable intervention on potentially high cost and avoidable hospital admissions. This is important as decreasing those admissions is seen as a way to improve health outcomes and control high cost patient expenditure (Billings, Dixon, Mijanovich & Wennberg, 2006).
One of the first models to come out of the NHS was by Billings et al. (2006). They developed a method to identify patients at high risk of hospital admission over a period of 12 months for use in primary care facilities and general practices (GPs). Their patients at risk for re-hospitalisation (PARR) algorithm focused on admissions for a number of "reference" conditions such as

congestive heart failure, chronic pulmonary disease and diabetes rather than all hospital admissions (as accidents are harder to predict). A 12 month period was used as a triggering admission period for each patient and the model used three years of historical data to predict whether an admission would occur in the following 12 months. The 21 strongest variables were used in the final model including age, sex, ethnicity, hospital, number of previous admissions and the presence (or absence) of multiple clinical conditions. A probability or *risk score* for each patient was calculated using logistic regression which is multiplied by 100 to generate a score between 1 and 100. A high probability score indicates a greater risk of admission in the following 12 months. They used a different 10% sample of the data each for training and testing. They claim that the two most important indicators in assessing performance is the sensitivity (how well the model predicts high risk patients) and 1 - the positive predictive value (PPV). 1-PPV measures the number of patients identified as at risk that do not experience a readmission. This is a way of measuring the cost effectiveness of the model as you want to avoid dissipation of the money spent on interventions. With a risk score threshold of 50 (probability of 0.5), the algorithm found that 34.7% of patients were incorrectly identified as at risk (a PPV score of 65.3%). The receiver operating characteristic (ROC) area under the curve measures the trade off between the sensitivity (true positive rate) and 1 - the specificity (the false positive rate). The value for this model was 0.685 indicating a 68.5% probability that a randomly selected patient with a future admission will receive a higher risk score than a randomly selected patient who will not have a future admission. They also performed business case analysis by assuming three different intervention costs (£500, £750 and £1000 per patient at risk) and three different hospital admission reductions (10%, 15% and 20%) at risk score thresholds of 50, 70 and 80. They found that for a risk score of above 70 (PPV 77.4%) this results in a saving of £750 or less. For risk threshold of 80 savings can be predicted across all assumptions.

Many other studies have developed 12 month predictive risk models similar to PARR such as one developed in Scotland by their National Health Services for patients greater than 65 years old (NHS National Services Scotland, 2011). Their logistic regression model resulted in a PPV of 53% at a threshold 0.5 and a ROC area under the curve value of 0.68.

Another similar study in the UK by Bottle, Aylin & Majeed (2006) identified acute inpatients at high risk of having at least two further future acute hospital admissions in the 12 months following the initial admission. A variable included in this study is the Charlson Comorbidity Index (CCI) which gives various weights to the presence of conditions such as diabetes and congestive heart failure based on diagnosis coding. They used a 50% training dataset for logistic regression models, 50% testing dataset for validation. The sensitivity, specificity, Positive Predictive Value (PPV) and area under the ROC curve were measured. Results show that an increasing risk threshold (therefore less patients flagged as at risk) decreases sensitivity (proportion of patients correctly identified by the mode over the total who have two or more admissions in the following 12 months), increasing specificity (correctly identified as not admitting) and increasing PPV (proportion of flagged patients correctly identified).

Choudhry et al. (2013) created 30 day hospital readmission models to identify high risk patients that require intervention resource in the US. This model relied on the ROC area under the curve statistic (C-statistic) as its main accuracy measure. They believe a value less than 0.6 has no clinical value, 0.6-0.7 has limited value, 0.7-0.8 has modest value and a C-statistic greater than 0.8 is adequate for clinical use. Models were developed using 1 year of data split into 75% derivation and 25% validation datasets. The model was fit using bootstrapping methods by randomly sampling two-thirds of the data into the derivation dataset 500 times and then averaging the coefficients which were

then applied to the derivation dataset. Predictor variables were included in the model if they had a p-value ≤ 0.05. Two models were developed, one for use when a patient is admitted to hospital and one post discharge. The admission model C-statistic was 0.76 and the discharge model 0.78 (which included variables only available after discharge such as length of stay and clinically coded procedures and discharges).

An important factor in whether a patient readmits or not according to many clinicians is the level of social support they have. For example for an older patient, whether they go to a rest home or not and the level of care that may be provided to them at said rest home (hospital level or basic care). Many predictive risk studies do not incorporate social support information into their models. Hasan et al. (2010) developed a model to predict 30 day hospital readmissions in the US and set out to identify the patient factors that are significantly associated with high risk of readmission for general medicine patients. Potential variables were acquired through patient interviews within 48 hours of admission as well as using administrative data. The different types of potential variables included demographics (age, ethnicity etc), social support (marital status, help at home, regular physician) and health information (CCI, a self reported health rating etc). Two thirds of the data was used for model training and the remaining data for testing. A logistic regression model was fitted to the training data and only variables with a p-value less than 0.05 were included in the final model. The final model included the type of insurance, marital status, whether they have a regular physician, CCI, a physicality index score, admissions in the past year and length of stay (LOS). The ROC C-statistic was 0.65 in the derivation and 0.61 in the validation datasets.
In another study highlighting the importance of social information by Hersh, Masoudi & Allen (2013) researched the post discharge environment for heart failure patients. They highlighted that clinical data may not fully quantify the potential causes of readmission. Unfortunately social support information is

difficult and expensive (in both time and cost) to capture in New Zealand but these studies do show the importance of these kinds of variables in predictive models.

4.2.2 Patients at Risk of Readmitting Within 30 Days (PARR-30)

The most recent model to come out of the NHS is the Patients at risk of readmitting within 30 days (PARR-30) model (Billings et al., 2012). This is a development on the PARR model described earlier except the readmission period has reduced from 12 months to 30 days as the likelihood of an unplanned readmission is more likely in the immediate post discharge period. The rate of readmissions over this shorter period of time of approximately one month has become an important outcome measure in the health sector in the United Kingdom (UK) as it is here in New Zealand. The government has also proposed that the NHS should not pay hospitals for emergency readmissions within 30 days of a planned elective admission.

The logistic regression model was developed using a 10% sample of a 12 month period of data generating an expected probability of readmission within 30 days for each patient. The variables included in the model were chosen based on ease of use for an application that could be used at the patients bedside. These included age, domicile deprivation, type of admission, acute admission in the last 30 days, number of acute discharges in the past 12 months, the presence of 11 major health conditions and hospital. Probabilities were split into 20 risk bands where Band 1 had the lowest chance of readmission and Band 20 patients had the highest risk. They found that for higher risk patients the actual readmission rates were higher but the number of patients in these bands decreased as the risk increased.

To test model accuracy they used PPV (proportion of patients correctly

identified by the model as at risk over the total at risk), sensitivity (correctly identified by the model as at risk over the total that actually readmit) and the specificity (correctly identified by the model as low risk over the total that do not readmit). These measures can be traded off against each other by adjusting the risk threshold, with an increasing risk threshold the sensitivity decreases and the specificity increases. The ROC C-statistic is also calculated which summarises the trade off between the true positives (sensitivity) and false negatives (1-specificity) at all possible risk thresholds. The model used a bootstrap evaluation method for training by repeatedly randomly drawing a large number of samples from the training dataset, fitting models to each sample and calculating the average for the performance measures over the samples. At a risk threshold of 0.5 the PPV was 59.2%, specificity was 99.5% (reasonably high), sensitivity was 5.4% (quite low) and are under the ROC curve was 0.70 (most predictive models range between 0.5 and 0.72).

For cost analysis purposes the average readmission cost of all patients in each risk band was calculated and, similar to the PARR 12 month model, they calculated the estimated savings at different readmission reduction levels (10%, 15% and 20%) for different thresholds. They found that the mean readmission cost for all low risk patients was small because a lower proportion of them readmitted. Alternatively an interesting finding was that also in the low risk bands the average cost of patients who readmitted only was lower than the patients in the high risk bands. At a 0.5 threshold 6395 out of 576868 were identified as at risk. If the mean readmission costs were £1088 per patient and if they assume the rate of readmissions will reduce by 10% through intervention then they would only have to spend £109 per patient to break even.

For the purpose of this thesis we have used the PARR-30 model for comparison of model performance measures.

4.2.3 New Zealand Studies

In New Zealand Rumball-Smith et al. (2013) developed a model to compare 30 day unplanned readmission rates between Maori and New Zealand European adult patients to highlight racial disparities in hospital care. New Zealand patient data over 6 years was sourced from the National Minimum Dataset (NMDS) for patients with specific surgical procedures. Logistic regression models were developed to investigate the association between ethnicity and unplanned readmissions. Odds ratios and confidence intervals were used to show differences between the variables.

Another study in New Zealand used logistic regression models to risk stratify the Waitamata District Health Board (DHB) patient population in Auckland (Vaithianathan et al., 2012). Using the methodology of the NHS PARR 12 month model they predicted readmissions to hospital within 12 months using all hospital admissions (unlike PARR which only included specific conditions). They used a 50% dataset each for training and testing. Significant variables included in the model were functions of sex, age, Diagnostic Related Group (DRG), number of admissions in the last 6 months, LOS and cost-weights. At four different risk thresholds (70, 80, 90 and 99) the total patients flagged as at risk, PPV, 1-PPV, sensitivity, specificity and the average number of readmissions for correctly flagged patients were calculated. A risk score threshold of 70 for this model resulted in a PPV of 73.37% and at a threshold of 90 the PPV was 91.67%. The ROC C-statistic was 71.18%. For their cost analysis they assumed three different costs of intervention \$500, \$750 and \$1000. Assuming the model will reduce readmission rates by 10% at a risk the savings were greatest for the lower intervention cost of \$500 at a risk threshold of 80.

Although the 12 month period between initial discharge and readmission is used in a few of the models discussed in this chapter we feel that that prolonged period of time may identify readmissions that are unrelated to the initial ad-

mission. It is for the reason we focus on the shorter period of approximately one month in this thesis.

Chapter 5

Data

5.1 Waikato District Health Board

Waikato District Health Board (DHB) covers the geographical area from part of the Ruapehu region in the south, right up to Coromandal peninsula in the north, Raglan on the west coast and over to Waihi on the east coast of New Zealand. The population was approximately 371,540 in the 2013 fiscal year. Waikato DHB is made up of a diverse population that fits into ten different territorial local authorities (TLAs), of which the city of Hamilton is the largest. This is where Waikato hospital, a major tertiary hospital in New Zealand, is based. It is the largest of the five main hospitals included in this analysis which also includes the four rural hospitals Taumarunui, Thames, Te Kuiti and Tokoroa as well as private hospitals where elective surgeries are performed.

5.2 Data

Data was obtained from the DHB under permission from the Chief Operating Officer (COO) of Health Waikato, Waikato DHB for the purpose of this thesis. The extraction and analysis of this data meets the University of Waikato ethics approval.

In New Zealand we are very fortunate to have a unique National Health Index

(NHI) number for all patients that have encountered our health system at any point in time. This means that we can link a patient's health experience over time through New Zealand health services. The NHI is encrypted in the DHB database for the purposes of this study so a patient is not identifiable in the dataset used in this analysis.

The data used in this model is from the Waikato DHB system CostPro which stores data from the patient management system iPM on a daily basis. This patient management system is where all admission, transferring, discharge, emergency, clinic and theatre information about a patient is recorded. In CostPro each new episode is given a unique number so that patients can be tracked through the system. An episode is the singular hospital event (which can be inpatient, emergency, outpatient etc) and one person (or NHI) can have many throughout a lifetime.

Waikato DHB inpatient service data is sent to the MoH and accumulated into a large dataset known as the National Minimum Dataset (NMDS) with all other New Zealand DHB's data. Information is sent back to DHB's about their patients that are treated in other DHB's, not in their own hospitals (for example Waikato DHB patients treated in Rotorua hospital, which is in the Lakes DHB).

The data included in this analysis includes Waikato DHB service data for Waikato DHB domiciled patients for both the initial and readmission episode. We also include data from the National Minimum Dataset (NMDS) for Waikato DHB domiciled patients for the readmission episode only. Including this data means we can pick up all possible readmissions for Waikato DHB patients but only use Waikato DHB service data for the initial admission as we cannot improve on admissions to other DHB hospitals.

5.2.1 MoH Framework

The data in the model is built around the MoH framework for their quarterly performance measure, OS8: Acute Readmissions to Hospital, in which they use data from the NMDS for all DHBs in New Zealand to compare readmission rates amongst DHBs.

The qualifications for how this measure is calculated is detailed in this paragraph.The initial admission must be an Inpatient case weighted MoH funded event. This excludes patients discharged under the palliative care specialty, patients from overseas and episodes where the initial admission ended in a patient's death. Also excluded are patients under the specialty emergency medicine with a LOS less than 24 hours as these are patients who are in the Emergency Department (ED) for a short period of time. For the purposes of this analysis we have excluded all non Waikato DHB domiciled patients from the initial episode. Similar to the ED patients excluded under MoH criteria we exclude those that stay less than 24 hours in our Acute Medical Unit (similar to the Emergency Department but for medical patients) are excluded from the initial episode. In terms of the readmission episode for the MoH, it must be an acute (unplanned) admission where the days between the initial episode discharge date and the readmission episode admit date are less than or equal to 28 days. Patients that are transferred between hospitals within a 24 hour time frame are excluded from the analysis as well as statistical (or funding) admissions (for discharges to different services of a hospital within 24 hours; for example an orthopaedic patient to a rehabilitation ward). Data was investigated to ensure the data is correctly coded for transfers between rural Waikato DHB hospitals and Waikato hospital to be sure the query is working correctly. Also excluded from the dataset are planned readmissions which are found using the diagnosis related group (DRG) which is an automatically generated code that combines a patients diagnosis, procedure and length of stay which therefore equates to a patients consumption of resources for that episode. There are certain DRG combinations in the initial and readmission

episodes that are defined by the MoH as planned ahead of time, for example an initial admission of acute leukemia without catastrophic consequences and then a readmission episode with the same DRG. Investigation went into whether these planned admissions were accurate or not, particularly for oncology who have a high readmission rate. Discussion with medical staff at Waikato DHB concluded that the DRG combinations for the planned initial and acute admissions are accurate and to all be included as unplanned acute readmissions. Specialties such as oncology who deal with serious illnesses such as Leukemia have patients that are often admitted acutely frequently due to the nature of their condition.

This data only includes the readmission where the admit date is the closest to the initial discharge date as some patients readmit multiple times within a 28 day time period and vice versa for the initial admission (only the closest initial admission is included for each readmission).

Also excluded from the initial admission are patients with incorrect or absent domicile codes (due to incorrect or absent addresses in the system).

The data includes a total of 4 years and 4 months of data with an initial episode discharge date between July 2009 and October 2014 as the readmission data is available until the end of November 2013. The total observations in the 28 day and 14 day datasets is 233,334 patient episodes.

5.2.2 Variables

The response variable in this study is a binomial variable {0,1}, describing whether the patient had an acute readmission within 28 days of the initial episode as described in the MoH framework. All of the explanatory variables contain information about the initial episode. Many variables were considered to be included in the model. Variables were included in the dataset because of clinical importance, inclusion in the predictive risk models discussed in the previous chapter or because they were suggested by medical staff or manage-

ment at Waikato DHB.

The possible explanatory variables in this thesis include:

1. Time variables Fiscal year, Season, Financial quarter and Month
2. Hospital and Hospital Group
3. Specialty Cluster and Specialty Cluster Group
4. LOS and LOS Group
5. Ethnicity Description and Ethnicity Group
6. Age at admission and Age group
7. Sex
8. TLA and TLA Group
9. Deprivation Score and Deprivation Score Group
10. Patient Category
11. Admit type
12. CCI weight and CCI Group
13. 20 disease variables: Cerebrovascular disease, Chronic pulmonary disease, Congestive heart failure, Connective tissue disease, Dementia, Diabetes, Mild liver disease, Myocardial infarction, Peripheral vascular disease, Any tumour, Hemiplegia, Leukaemia, Lymphoma, Moderate or severe renal disease, Ulcer disease, Moderate or severe liver disease, AIDS/HIV, Metastatic solid tumour, Delirium, Diabetes with end-organ damage
14. Number of Acute Admissions in the previous 12, 24 and 36 months and Groups

15. Number of Total Admissions in the previous 12, 24 and 36 months and Groups

16. Number of ED Presentations in the previous 12, 24 and 36 months and Groups

Included in the data are the four full fiscal years and one quarter of one year (2010-2014) (Table 5.1). A fiscal year runs from July to June. Seasons, financial quarters and months are also included as categorical variables in the analysis.

Table 5.1: 28 day readmission dataset readmission rates by Fiscal Year

Fiscal year	Readmission rate
2010	8.3%
2011	8.3%
2012	8.7%
2013	9.4%
2014	9.6%

The hospital variable includes six different hospitals Waikato, Taumarunui, Tokoroa, Thames, Te Kuiti and Private hospitals (this category is all initial admissions to private hospitals eg: elective surgeries under contracts with private hospitals). Another variable Hospital Group was created to cluster hospitals with similar readmission rates (Table 5.2).

The specialty cluster is made up of discharge health specialties grouped into planning clusters. The discharge specialty is the specialty under which the patient is discharged which reflects the nature of the services that is provided to that patient. These specialties have been grouped into eight specialty clusters: Women's health, all Surgical specialties, Orthopaedics, Paediatrics; CCTVS (Cardiology, Cardiovascular, Vascular and Thoracic Surgery),

Table 5.2: 28 day readmission dataset readmission rates by Hospital

Hospital Group	Hospital	Readmission rate
0	Private	3.7%
1	Waikato	8.6%
	Te Kuiti	8.7%
2	Thames	11.1%
	Tokoroa	13.9%
	Taumarunui	15.0%

all Medical specialties, Oncology and Emergency Medicine which are formulated for planning and management. For the purposes of this analysis another variable was created to combine clusters into groups with similar readmission rates (Group 1 includes Women's, surgery, Orthopaedics, Paediatrics; group 2 includes CCTVS and Medicine and group 3 includes Oncology and Emergency Medicine), see Table 5.3.

Table 5.3: 28 day readmission dataset readmission rates by Specialty Cluster

Cluster Group	Specialty Cluster	Readmission rate
1	Womans Health	4.1%
	Paediatrics	5.6%
	Orthopaedics	5.7%
	Surgery	7.4%
2	CCTV	12.8%
	Internal Medicine	13.3%
3	Oncology	16.4%
	Emergency Medicine	17.5%

Length of stay (LOS) is the time between admit date and time to hospital

Table 5.4: 28 day readmission dataset readmission rates by LOS

LOS Group	Readmission rate
1	6.1%
2	11.1%
3	14.1%
4	15.3%

and date and time of discharge. The hospital average length of stay is currently approximately 5 days. For the purposes of this analysis we tried LOS as a continuous variable and a categorical with four groups (1: LOS less than 2 days, 2: LOS less than 5 days, 3: LOS less than 10 days and 4: LOS greater than 10 days). See Table 5.4 for the readmission rates.

Table 5.5: 28 day readmission dataset readmission rates by Sex

Sex	Readmission rate
Female	8.3%
Male	9.3%

Three demographic variables are included. Ethnicity Description which is at a low level (for example African, Cook Island Maori, Chinese, Latin American) or Ethnicity Group which is a higher planning level of just four ethnic groups (Other, Maori, Pacific Islander or Asian). We tried age as both a continuous variable and a categorical variable with four groups (0-14, 15-39, 40-64 and 65 or more years). Age is calculated as the years between date of birth to admit date of initial admission. Sex is also included (Table 5.5).

TLA was also included as a variable in the model. There are 10 different TLAs in the Waikato DHB region. We included that as a variable and also

Table 5.6: 28 day readmission dataset readmission rates by TLA

TLA Group	TLA	Readmission rate
0	Matamata-Piako	7.3%
	Otorohanga	7.7%
	Waikato	8.0%
	Waipa	8.2%
	Thames-Coromandel	8.7%
	Waitomo	8.7%
1	Hamilton	8.9%
2	South Waikato	9.3%
	Hauraki	10.4%
	Ruapehu	12.0%

in similar readmission rate groups (Table 5.6) . Related to the TLA variable is the deprivation score. This is an index of socioeconomic deprivation created by Statistics New Zealand based on domiciles (addresses). This variable gives a number from 1 to 10 on an ordinal scale where 1 is least deprived and 10 is most deprived. We used this as a categorical variable with 10 categories as well as grouping it into similar clusters 1-3, 4-6 and 7-10 (Table 5.7).

An *inpatient* event is where a patient is admitted to hospital with a patient type of either daycase (a MoH defined event when a patients LOS is 0 at midnight because they are admitted and discharged on the same day) or an inpatient (where all other patients are inpatients). This event can be an acute (an unplanned admission), an arranged admission (where a patient is booked to come into a facility within approximately 7 days, these are not Emergency Department or waitlist patients) or an elective (a planned admission from the waiting list or booking list) admission.

Table 5.7: 28 day readmission dataset readmission rates by Deprivation Score

Deprivation Group	Deprivation Score	Readmission rate
1-3	1	7.4%
	2	7.3%
	3	7.6%
4-6	4	8.3%
	5	8.3%
	6	9.1%
7-10	7	7.9%
	8	9.1%
	9	9.8%
	10	9.8%

Hospital clinical coders use the 10th version of the International Statistical Classification of Diseases and Related Health Problems (ICD-10) 6th edition to code patient diagnosis which is a patients reason for being in hospital as well as other comorbidities. Upon discharge from hospital patients notes are sent to the clinical coders who then code all diagnoses relating to that discharge or that effected the patient during their hospital event. These diagnoses are ordered where the most prominent reason for being in hospital is coded as the primary diagnosis and so on.

Many articles and medical staff at Waikato DHB mention that the CCI is a good way to summarise an individual patients disease comorbidities. This index was developed in 1987 using data from a medical patients and validated on breast cancer patients (Sarfati, Tan, Blakely & Pearce, 2011) as a way to quantify patient comorbidities. In this analysis the diseases used in this variable are found from the ICD-10 coding based on the coding in the (Gabbe, Harrison, Lyons, Edwards & Cameron, 2013) article with a few minor adjustments to the ICD-10 codes as validated by the Waikato DHB clinical coders.

Table 5.8: 28 day readmission dataset readmission rates by CCI weight group

CCI Group	Readmission rate
0	6.7%
1	13.7%
2	12.3%
3+	15.0%

All possible coded diseases are used in this analysis (no matter the order of importance). This measure assigns a weight to each disease that a patient has which results in an overall score. A score of 0 indicates that none of the conditions are present and a high score indicate many are present therefore, a high level of comorbidity. There are 19 variables included in this index. Worth a weight of 1 are: Myocardial infarction, Congestive heart failure, Peripheral vascular disease, Dementia, Cerebrovascular disease, Chronic pulmonary disease, Connective tissue disease, Ulcer disease, Mild liver disease and Diabetes. Hemiplegia, Moderate or severe renal disease, Diabetes with end-organ damage, Any tumour, Leukaemia and Lymphoma have a weighting of 2. Moderate or severe liver disease has a weight of 3 and the remaining diseases, Metastatic solid tumour and AIDS/HIV, are the most severe with a weight of 6. In the model we used CCI as both a continuous variable and a categorical variable with 6 groups (0, 1, 2, 3 or 4 or 5+ for a weighting of 5 or more), see Table 5.8.

We took the 19 diseases that were included in the CCI variable as well as delirium (upon speaking with a Waikato DHB medical staff member he deemed it as an important variable) and used each of them as binomial variables for disease presence or absence. An important note about delirium is that it is speculated to be not well documented in patient notes therefore the number of patients that have delirium may not be fully represented in this variable.

International studies also include the number of acute admissions in the last 12, 24 and 36 months, these are unplanned emergency admissions to hospital. We also considered the number of total admissions (including acute, arranged and elective) in the last 12, 24 and 36 months. A different variable we also included was the number of presentations to any Waikato DHB Emergency departments in the 12, 24 and 36 months previous to the initial admission. All of these variables mimic those used in the PARR-30 model. We treated these variables in the dataset as both continuous variables by counting the totals as well as each as a group variables (0,1,2,3,4 and 5+) to see which performed the best.

5.3 Days Between Initial Admission and Readmission

As mentioned previously in this thesis we found that the period of approximately 1 month is the most commonly used readmission period in both international and New Zealand readmission analysis. We could not find any scientific reasoning behind the use of this time period so we decided to compare longer (and one shorter) time periods to see what the effect would be on our model performance measures and cost analysis over varying risk thresholds. We did this by varying the number of days between the initial admission discharge date and the readmission date. We experimented with 14 days (2 weeks), 42 days (6 weeks), 56 days (8 weeks), 82 days (3 months approximately), 182 days (approximately 6 months) and 365 days (1 year). These different time frames were chosen to investigate the effect on the prediction of the model and model measures. We could not use the same data for all of these time periods as we need to reduce the initial date range to allow time for readmissions to occur. The 14 day analysis date range remained the same as the 28 day data. For the

6 and 8 week datasets the date range reduced to ending on 30 September 2013 with 228,592 observations. The 84 day data date range reduced to the end of August 2013 with 223,840 observations. For 182 days we reduced the date range to 31 May 2013 with 209,376 observations. Finally, for 365 days the date range reduced to end in November 2012 using 182,054 observations.For the 56 day readmission dataset we also calculated whether the readmission occurred within 28 days from the initial admission or not. This was to see how many 28 day readmissions were predicted in each risk band.

5.4 Actual Total Cost of Readmission

This thesis also sets out to perform a cost analysis to compare number of days for readmission period, risk threshold and model selection. For this analysis we need to calculate the actual total cost of each readmission episode. Because we know the readmission episode occurred then we know how much money this episode cost the DHB. This cost is from the DHB costing system which uses activity based costing meaning all activity related to the patient episodes actual ward, radiology, theatre costs as well as Doctor and Nurse costs (based on averages around doctor and nurse time over large groups) are included in this amount. If a patient does have an acute readmission then the cost of readmission is $0.

Chapter 6

Data Analysis

6.1 Modelling Data

This thesis deals with probabilistic classification methods. The two important methods used are logistic regression and Naive Bayes models. The purpose of this study is to compare the two methods and decide what one performs better given its use in predictive risk modelling.

6.1.1 Logistic Regression

Logistic regression models were fit to the data in the statistical computing program R using the built in GLM function using the the binary attribute, readmission, as the response variable and the logit function as the link function. The reference condition for each variable included in the analysis was the category with the lowest readmission rate observed in the data.

Summary statistics for each of the variables tested in the model were calculated. This included the coefficient, odds ratio (OR) and coefficient p-values. The residual deviances and the Akaike Information Criterion (AIC) were used to as goodness of fit statistics for model comparison. The odds ratios (ORs) for the coefficients were calculated by taking the exponential of the variable coefficients. The OR describes the increasing or decreasing odds of readmission

for each of the explanatory variables and their levels in the model. To decide if explanatory variables are valuable in the model the ORs are compared with the actual readmission rates to ensure the variables are sensible.

6.1.2 Bayesian Belief Networks

The specific type of Bayesian network we fit to the data is the Naive Bayes (NB) model. This was done also in R using the naive.bayes function in the bnlearn package. Prior probabilities are calculated based on the class variable readmission. The conditional probability for each explanatory variable is then estimated using the Maximum A Posteriori (MAP) as described in Chapter 3 of this thesis.

6.2 Model Performance Measures

Various measures were used to measure model performance in line with what many studies described earlier in this thesis do. These measures were calculated in cross validation using a risk threshold of 0.5 and averaged across the folds. They were also calculated in the risk band tables at each risk threshold. This made it simple to compare different risk cut offs and different models (GLM versus NB) and different readmission day periods.

6.2.1 Positive Predictive Values

The positive predictive value (PPV) is the proportion of the patients identified by the model as at risk who experience a readmission over the total patients identified by the model as at risk. A high PPV value will indicate that the high proportion of patients that do experience a costly readmission which is likely to be something hospitals want to prevent. A low PPV value would mean that many patients that were identified as high risk did not readmit therefore the intervention would be wasted (Lewis et al., 2011).

Similarly the Negative predictive Value (NPV) is the proportion patients correctly identified as low risk over the total patients that are not at risk. Generally this value is high as the model tends to correctly identify patients that are low risk better than it correctly identifies patients at high risk. A high value also means that potential intervention costs are not being wasted because the false negatives are low. This value is often large in risk predictive models as the total patients who do not readmit is always a lot higher than the total that do readmit

6.2.2 Sensitivity

The sensitivity is the percentage of those who are correctly identified by the model as at risk over the total number that actually readmit. The sensitivity is also known as the true positive rate. Unfortunately this measure tends to mask the potential value of models in targeting preventive interventions (Billings et al., 2012). Rather than measuring the how well the model is correctly predicting those at risk as the PPV does, it looks at the total correctly identified over the total who readmit which is always going to be high for low risk bands because you are predicting basically everyone at risk. Alternatively at high risk bands you are predicting few patients at risk therefore the sensitivity is quite low.

1 - sensitivity or the false negative rate is a good measure for looking at the time period between initial admission discharge and readmission. If the time period is too short (ie 14 days) then the false negative rate is too high. This is because there are less patients that actually readmit in that shorter time frame.

6.2.3 Specificity

The specificity predicts which patients will not have a future acute readmission. It is the proportion of low risk patients that do not readmit divided by the total who do not readmit. It is also known as the true negative rate. The

specificity is not always regarded as a useful measure as the majority of the patients in risk predictive modelling do not readmit (the actual readmission percentage is usually less than approximately 20%) therefore the rate will always be reasonably high.

1-specificity, like the false negative rate, is also good for comparing time between initial discharge and readmissions. If the time period is too long (ie 56 days) then the are too many false positives because there are less patients in the dataset that do not readmit (therefore increasing the proportion). This is also a good measure for cost analysis as a high false positive rate indicates wasted intervention costs on patients that do not readmit even though they are identified as at risk.

6.2.4 Receiver Operating Characteristic Curve

The receiver operating characteristic (ROC) curve models the trade off between the true positive rate (sensitivity) and the false positive rate (1 - the specificity). These values are plotted against each other at all risk cut off levels. The area under the curve, also known as the curve statistic (C-statistic), is a value that lies between 0 and 1 so you can compare the sensitivities and specificities between different models. If a model is performing better than a random guess then the C-statistic value will be greater than 0.5.

6.2.5 Summary of Performance Measures

The PPV is a good measure for testing model performance as it calculates the number of correctly identified patients in each risk band. Alternatively the sensitivity calculates the number of true positives out of the total that readmit. This does not measure whether the model is correctly identifying patients like PPV, rather it measures whether you are identifying enough patients compared tot he true readmission rate. Similar to that the specificity measures whether

you are correctly predicting enough patients as low risk compared to the true non-readmission rate. A better measure for assessing the models identification of low risk patients is NPV as this is the total correctly identified as low risk over the total at low risk. Again measuring the performance of the model at each risk threshold, not just the total identified. Therefore in this theses we regard the PPV and as the most relevant model performance measure.

6.3 Cross Validation

To test the performance of an algorithm on our data we need to assess the error rate on a dataset that had nothing to do with the formation of the model. The data used for the model formulation is called the training dataset and the data used to assess that model is known as the test dataset (Witten & Frank, 2005). We assume that the training and test datasets are representative samples of the underlying population. Witten & Frank (2005) believe that the larger the training sample the better the model therefore we choose a small test dataset. 10-fold cross-validation splits the data into 10 groups where 9/10 of the data is used of training and 1/10 used for testing. This is then repeated 10 times until each instance has been used for testing once. The error rate and performance measures are calculated each time on the test dataset then averaged across all 10 holdout sets to get an overall error estimates and measures. Performing cross validation on the dataset results in an error rate with a small standard deviation which then decreases as the validation is repeated ten times. Using 10 folds is the standard method as it results in the best error estimate.

The 10-fold cross validation we used in our analysis is based on the cv.glm function in R. Every time the function is run each NHI is randomly allocated into 1 of 10 folds. Each NHI (or person) may have multiple episodes so each fold does not have exactly the same number of episodes but it is approximately even. We then run model on 9/10 of data and test using predict on the other

1/10 and repeat for each fold. Using a risk threshold of 0.5 we calculate the PPV, sensitivity, specificity and ROC curve statistic at each fold as well as the total in each fold that are predicted at risk. Those values are saved and then averaged at the end to get final performance measures from the cross validation. We then compared these results to the model trained and tested on the whole dataset and we get similar results.

6.4 Cost Analysis

Cost analysis is a way to measure the potential savings of intervention on patients identified by the model as high risk. We calculated the average cost of a readmission at each risk band for both the total patients in each band and for the total that actually readmitted similar to the PARR-30 model by (Billings et al., 2012). We used the cost described in the previous chapter which is the actual total cost of the readmission episode. We can analyse the potential savings at each risk threshold to estimate the potential savings to the DHB from a reduction in readmissions.

The cost analysis is performed at each possible threshold.

The total savings from using the model is calculated using the following formula:

$$\text{Net savings} = \text{Cost of true positives} - (\text{Intervention cost} \times \text{Total at risk})$$

Where the cost of the intervention is approximated at three levels: $500 (which is the approximate cost of a patient in the Waikato visiting a General Practitioner once a month for one year), $1000 (which is the rough approximate cost of one bed day at the Waikato DHB including ward stay and diagnostics) and $2000 (included as a possible high cost intervention). The total cost of intervention is the total at risk multiplied the intervention cost. The total patients at risk is the number of patients in each threshold that are deemed at risk if they have a probability greater than or equal to that risk threshold. The

true positives are the number of correctly identified at risk patients. The cost of those patients is the average readmission cost for all at risk patients that do readmit multiplied by the number of true positives. The potential savings for the DHB is calculated by subtracting the cost of intervention on at risk patients from the true positives. The potential savings can be found at each risk threshold for different intervention costs.

The cost discussed previously is based around the assumption that all of the patients that are predicted at risk do not readmit to hospital. However, this is an unlikely assumption as we cannot guarantee all at risk patients do not return. Vaithianathan et al. (2012) performed a Business Case analysis assuming different reductions in admissions. For the second part of the cost analysis different readmission reduction rates are compared to the assumption that 100% of patients will not readmit. To make the cost analysis more realistic different readmission reduction rates are used; a pessimistic 10% , 20% and an optimistic reduction of 50%. To get the reduced reduction numbers the true positives are multiplied by the new reduction levels. To get the cost saved from those patients not returning that number would then be multiplied by the average readmission cost of the true positives at each risk threshold. This is calculated as the total readmission cost for the true positives divided by the total true positives. The cost of the interventions remains as it is in the previous cost analysis. This second cost analysis is performed on the 56 day model only.

6.5 Risk Band Table

6.5.1 Risk Threshold

The *risk threshold* is the probability cut off point at which the number of patients predicted to be at risk, and alternatively not at risk. The most common risk threshold in the articles described previously is 0.5. However a few models

do use higher levels of 0.7 and above. Using different risk thresholds result in varying performance measures as you are altering the number of patients at risk. For example for a measure such as sensitivity an increasing risk threshold will often mean a smaller sensitivity as you are predicting less patients at risk and therefore less correctly identified as at risk over the total readmissions.

6.5.2 Risk Band Table

Risk band tables are created for this study, similar to that in the PARR-30 model (Billings et al., 2012). This summarises the output of the models into 20 probability bands each including a probability of 0.05. To create the risk band tables the probabilities were calculated for each observation using the complete dataset for both the training of the model and testing of the model. Then in each of the 20 bands the following was calculated: total in each band, percent of the total in the analysis, the number of patients that do actually readmit, the number that do not readmit, the proportion of patients that readmit, the total cost of readmissions, the average cost of readmission for the total in each band and the average cost of readmissions for patients that readmit only. For the 56 day readmission data only the total 28 day readmissions in each risk band were also calculated to observe the total 28 day readmissions at different thresholds in the 56 days model compared to the 28 days model.

Using those summaries at each risk band level the risk threshold measures can be calculated. These are the total at risk, total not at risk, true positives, false positives, true negatives and false negatives. These values are used to calculate the model performance measures at each risk threshold: PPV, NPV, sensitivity and specificity. Using the sensitivity and 1 - the specificity the ROC curve is plotted also. The performance measures in the risk tables were compared to the results from the 10 fold cross validation procedure. The cross validation averages for PPV, sensitivity and specificity were very similar to

the results in the risk band tables.

Chapter 7

Results

In this chapter we describe the different criteria considered to find the optimal model for predicting patients at risk of readmission. These include

1. Statistical criteria such as coefficient p-values, odds ratios, AIC and residual deviance, performance measures such as PPV, sensitivity etc to find the best model.

2. Predictive criteria such as the optimal number of days between the initial discharge and readmission and risk threshold level.

3. Cost analysis by using the actual cost of readmissions to calculate the possible savings if the model is used

In this thesis we set out to clarify what the ideal strategy is that saves the DHB the most money considering each of the criteria above. The predictive criteria above effects the number of patients that at identified at risk and therefore the total intervention cost and the potential savings from model utilisation. The cost analysis and the PPV are the main methods used to find the best risk threshold and days between initial discharge and readmission. Although the MoH reports focus on the 28 day period this thesis sets out to test whether that this is, in fact, the optimal time period and if not, what period should we focus on for predictive modelling purposes.

7.1 Model Selection

The analysis of the data in this study starts with finding the statistically optimal model in terms of explanatory variable selection and model fit. To do this the model goodness of fit statistics AIC and Residual Deviance are used as well the coefficient p-values and odds ratios. The cross validation performance measures PPV, sensitivity, specificity and ROC C-statistic are also used to find the optimal model.

Logistic regression models are fit to the 28 day readmission data R based on the theory discussed in the second chapter of this thesis using the binary response variable $\{0, 1\}$ where 0 means no readmission and 1 is a readmission. The summary output from these models are the coefficients, their odds ratios and their p-values. The model AIC and the residual deviances are also calculated. These are calculated by fitting the model on the whole dataset and using the whole dataset for prediction to get the fitted values. 10-fold cross validation was performed on each of the models to validate the performance measures at a risk threshold of 0.5. Using this we find the average PPV, sensitivity, specificity and the ROC C-Statistic. The sum of the total number at risk per fold was calculated to approximate the number of patients at risk for each model. In each of the models we checked that the cross-validation statistics are more or less equal to the statistics calculated using the whole dataset for training and testing.
The explanatory variables were used in the logistic regression models as single explanatory variables, including and excluding the intercept term, to check the results of the multiple logistic regression below are sensible.

Many different models were tested with different combinations of explanatory variables to keep it brief five different models are discussed from different stages of the model selection process. This includes the reasons for including and excluding different variables. The output for each can be found in Table

Table 7.1: Generalised Linear Model Selection

Model	Res Dev	AIC	At risk	Sensitivity	Specificity	PPV	ROC
GLM 1	123991	124179	428	0.8%	99.9%	38.0%	0.742
GLM 2	124265	124369	400	0.8%	99.9%	39.4%	0.740
GLM 3	124274	124372	405	0.8%	99.9%	38.6%	0.740
GLM 4	124630	124708	456	0.9%	99.9%	40.1%	0.738

7.1. The models are described below:

1. The first model is a reflection of the start of the model selection process with nearly all of the variables in the dataset included. All continuous variables were removed from this model in favour of categorical variables. This meant variables such as the number of acute readmissions in the last 12 months were replaced by the same variable but using six categories to group the count as described in the data explanation Chapter 5. Other continuous variables excluded in this model in favour of categorical variables were age, LOS, CCI weight, TLA, hospital, ethnicity, deprivation score, specialty clusters and the count of admissions and ED presentation variables. For variables such as ethnicity we opted for the variable with the smallest number of categories (4 versus 26 levels).

   The results from this model can be found in Table 7.1 in this first row. Some variables included in the first model were not significant so were removed from the model. These were the 24 month admissions, the total admissions for 12, 24 and 36 month periods, hospital group, sex, the diseases Any tumour, Ulcer disease and Delirium.

   The coefficient odds ratios (ORs) were compared to the true readmission rates for each of the variables. For a few of variables the ORs were not comparable to the actual readmission percentages for some of the levels within the variable. An example is the TLA group variable where the

acutal readmission rate for the level 0 TLA is 8.1%, 8.9% for level 1 and the highest readmission rate is for level 2 with 10.3% readmitting. However the OR for level 1 was 1.15 and for level 2 the OR is 0.98. The odds of readmitting for patients in the level 2 group for TLA should be greater than the level 1 group as the readmission rate is. However it is not and it is for this reason TLA group and the variables ethnicity group, deprivation score and a few of the disease variables (Cerebrovascular disease, Connective tissue disease, Dementia, Hemiplegia) were removed from this model.

2. For the second model the variables described above were removed. The model summary statistics in Table 7.1 row 2 did not change much although compared to the first model. The PPV does increase by a 1.4%.

3. For the third model the purpose was to test whether the CCI group variable by itself or each of the remaining disease variables on their own have greater predictive power the our model. Each of the two were excluded from the model separately, the results from excluding the CCI variable are included in the table. The summary statistics were better for the model excluding the CCI variable it was exclude it from the model and include the remaining significant disease variables. However a few disease variables were included in the model at this stage even though their p-values were not significant. These were diabetes, mild liver disease, peripheral vascular disease and diabetes with end organ damage. They were included because of the clinical significance that they hold.

4. The final stage excluded the 36 month acute admission and ED presentations from the model compared only including the same variables over a period of 24 months. The model had greater power when the 36 month variables were removed from the model as the number of patients predicted at risk decreased to only 65 versus 456 in the 12 month model. All

the other summary statistics decreased also. It is for that reason the 36 month acute admissions and ED presentations were excluded from the final model.

7.1.1 Final model

The best model in terms of variable selection and model fit includes the following 21 variables

1. The fiscal year of the initial admission
2. The discharge specialty cluster group for the initial admission
3. The length of stay (LOS) group (0-1, 2-5, 5-10 and 10+ days) for the initial admission
4. Age group for age at the time of initial admission
5. The patient category (Inpatient or Daycase) for the initial admission
6. The initial admission admit type (acute, arranged or elective)
7. The presence of 13 possible diseases drawn from the CCI variable for the initial admission. This included (all as separate variables) Myocardial infarction, Congestive heart failure, Peripheral vascular disease, Chronic pulmonary disease, Mild liver disease, Diabetes, Moderate or severe renal disease, Diabetes with end-organ damage, Leukaemia, Lymphoma, Moderate or severe liver disease, Metastatic solid tumour and AIDS/HIV.
8. The number of acute admissions in the previous 12 months from the initial admission grouped variable
9. The number of ED presentations in the previous 12 months from the initial admission grouped variable

These variables were discussed with Waikato DHB hospital management and medical staff to confirm their clinical significance. The total at risk, AIC,

residual deviance, PPV, sensitivity, specificity and ROC values can be seen in the final row of Table 7.1. This model predicted the highest number of patients at risk as well as the highest PPV among all of the models. This means that although the model is predicting a high number of patients at risk, 456, it is also predicting them in those high risk bands 40.1% of the time correctly. The PPV is significantly lower than that predicted in the PARR-30 model Billings et al. (2012), 59.2%. The specificity of 99.9% is only slightly greater than PARR-30 99.5% and the sensitivity in this model is very low 0.9% compared to PARR-30 5.4%. The ROC C-Statistic, 0.74, is similar to PARR-30 0.70.

A risk band table was created as described in the previous section to compare to the PARR-30 model. The results of the 28 day final model did not predict patients in the high risk bands as the PARR-30 study did. This could be an indication that the 28 day readmission period does not allow time for patients to readmit. This is demonstrated in Section (7.2.5) where the model is run on the long time period datasets 84, 182 and 365. The risk band Table (7.4) shows that the model predicts patients in the high risk bands, this is discussed further in Section (7.2.5).

To test that there was nothing wrong with the final model artificial data was generated by using the high readmission rate categories for each of the variables from the final model. Approximately 2652 artificial patients were tested with "worst case possible" variables (for example 65+ age group and 5+ acute admissions in the last 12 months) and found that they did have probability values in the high risk bands. This confirmed that the model does predict correctly, the patients in the actual dataset simply do not fit into those extreme categories.

7.2 Final model analysis

In this section the final model, including the explanatory variables described in the previous section, is used to analyse different initial discharge to readmission time periods and different risk thresholds. This is to test whether the time period of 28 days has more predictive power over shorter and longer periods. The risk threshold of 0.5 is used in many of the models discussed in this thesis therefore, manipulating this threshold is also analysed in the following risk band tables so the best predictive criteria can be found.

Table 7.2: Logistic Regression Model risk band table 28 days

Risk Band	Probability	Total at risk	% of total	Total readmit	Total do not readmit	% actually readmit	Total cost of readmission	Average cost of readmission
1	(0.00-0.05)	85939	36.83%	2215	83724	2.6%	$8592278	$3879
2	(0.05-0.10)	82458	35.34%	6328	76130	7.7%	$41874660	$6617
3	(0.10-0.15)	30749	13.18%	4020	26729	13.1%	$29381500	$7309
4	(0.15-0.20)	15038	6.44%	2662	12376	17.7%	$19646000	$7380
5	(0.20-0.25)	8164	3.5%	1851	6313	22.7%	$13477340	$7281
6	(0.25-0.30)	4628	1.98%	1218	3410	26.3%	$9523465	$7819
7	(0.30-0.35)	2556	1.1%	742	1814	29%	$5342756	$7200
8	(0.35-0.40)	1610	0.69%	551	1059	34.2%	$3727887	$6766
9	(0.40-0.45)	1010	0.43%	350	660	34.7%	$2293808	$6554
10	(0.45-0.50)	726	0.31%	274	452	37.7%	$2023949	$7387
11	(0.50-0.55)	313	0.13%	124	189	39.6%	$1199639	$9675
12	(0.55-0.60)	112	0.05%	44	68	39.3%	$416881	$9475
13	(0.60-0.65)	27	0.01%	13	14	48.1%	$147444	$11342
14	(0.65-0.70)	4	0.002%	2	2	50%	$32804	$16402
15	(0.70-0.75)							
16	(0.75-0.80)							
17	(0.80-0.85)							
18	(0.85-0.90)							
19	(0.90-0.95)							
20	(0.95-1.00)							
	Total	233334	100%	20394	212940	8.7%	$137680411	$6751

In Table 7.2 the results for the 28 day final model are presented. The patterns in all of the measures, Total at risk, readmission totals and costs, can be seen across all of the readmission time periods. The actual readmission percent in each band increases as the probability increases but the total in each band decreases as the probability increases. This indicates that although there is a higher proportion of patients that do actually readmit in the higher risk bands, the total in those bands are only a small share of the total patients analysed. Similarly the same patterns are seen amongst the different performance measures. As the risk threshold increases the positive predictive value increases, the negative predictive value decreases slightly, the sensitivity rapidly decreases, 1 - specificity also rapidly decreases, the specificity rapidly increases and the same is seen for 1- sensitivity which rapidly increases.

Tables 7.3 and 7.4 compare the number of patients at risk and the PPV at each risk threshold for the 14, 28, 42, 56, 84, 182 and 365 day models. An important thing to note in these tables is that the is that there basically no one at risk of readmission in the 14 day model (only 5 at a risk threshold of 0.5). As the days between initial discharge and readmission increase, the number in the high risk bands also increases. For the 365 day model there is a significant number of patients in the high risk bands as the likelihood of patients readmitting is high in this long time period. Another important note to take away from these tables is that the PPV increases not only as the risk threshold increases but also as the readmission peirod increases. This indicates that the longer time period models are proportionally predicting a higher number of patients correctly than the shorter time periods (see Figure 7.1).

Table 7.3: Logistic Regression Model total at risk and PPV by risk threshold and 14, 28, 42 and 56 readmission days

Risk	Probability	14 days		28days		42 days		56 days	
Band		At risk	PPV	At risk	PPV	At risk	PPV	At risk	PPV
1	(0.00-0.05)	233330	5.9%	233334	8.7%	228592	10.7%	228592	12.3%
2	(0.05-0.10)	104903	9.9%	147395	12.3%	161544	14.1%	174476	15.2%
3	(0.10-0.15)	34494	15.0%	64937	18.2%	84042	20.0%	98800	21.1%
4	(0.15-0.20)	13802	18.9%	34188	22.9%	47295	25.1%	57646	26.7%
5	(0.20-0.25)	6074	22.2%	19150	27.0%	29451	29.3%	37096	31.4%
6	(0.25-0.30)	2856	23.4%	10986	30.2%	18565	33.2%	24956	35.0%
7	(0.30-0.35)	1125	25.3%	6358	33.0%	11743	36.4%	16818	38.5%
8	(0.35-0.40)	386	29.0%	3802	35.7%	7407	38.8%	11197	41.5%
9	(0.40-0.45)	107	34.6%	2192	36.8%	4604	41.6%	7403	44.1%
10	(0.45-0.50)	27	44.4%	1182	38.7%	2835	43.3%	4808	46.3%
11	(0.50-0.55)	5	60.0%	456	40.1%	1680	45.4%	2932	48.4%
12	(0.55-0.60)			143	41.3%	939	47.9%	1755	50.0%
13	(0.60-0.65)			31	48.4%	311	50.5%	917	52.8%
14	(0.65-0.70)			4	50.0%	74	48.6%	317	55.2%
15	(0.70-0.75)					11	63.6%	59	50.8%
16	(0.75-0.80)					1	100.0%	4	75.0%
17	(0.80-0.85)								
18	(0.85-0.90)								
19	(0.90-0.95)								
20	(0.95-1.00)								

Table 7.4: Logistic Regression Model total at risk and PPV by risk threshold for 84, 182 and 365 readmission days

Risk	Probability	84 days		182 days		365 days	
Band		At risk	PPV	At risk	PPV	At risk	PPV
1	(0.00-0.05)	223840	14.7%	209376	19.8%	182054	25.7%
2	(0.05-0.10)	191433	16.6%	206432	20.1%	182054	25.7%
3	(0.10-0.15)	117516	22.7%	148859	25.2%	166089	27.4%
4	(0.15-0.20)	74674	28.2%	104410	30.8%	126468	32.4%
5	(0.20-0.25)	48645	33.3%	73271	36.2%	93871	37.8%
6	(0.25-0.30)	33957	37.6%	52725	41%	70380	42.6%
7	(0.30-0.35)	24313	41%	39385	45%	54712	46.5%
8	(0.35-0.40)	17143	44.4%	30044	48.4%	41455	50.5%
9	(0.40-0.45)	12064	47.4%	22558	51.5%	31593	54.2%
10	(0.45-0.50)	8373	50.5%	16839	54.5%	24409	57.1%
11	(0.50-0.55)	5653	52.2%	12180	57.5%	18252	59.9%
12	(0.55-0.60)	3616	54.4%	8666	60.3%	13717	62.6%
13	(0.60-0.65)	2210	56.1%	5964	62.7%	9420	65.6%
14	(0.65-0.70)	1209	58.6%	3721	64.3%	6254	67.8%
15	(0.70-0.75)	516	64.5%	2089	67.3%	3806	69.7%
16	(0.75-0.80)	92	62%	1023	70.7%	2047	72.8%
17	(0.80-0.85)	9	66.7%	378	77%	857	76.4%
18	(0.85-0.90)			38	76.3%	219	78.1%
19	(0.90-0.95)					16	75%
20	(0.95-1.00)						

7.2.1 14 days

The total at risk and the PPV can be seen it Table 7.3. This readmission period is does not predict many patients in the high risk bands (only 5 at a risk threshold of 0.5). The PPV is constantly lower than the other periods as seen in Figure 7.1. Except in the high risk thresholds where although it is

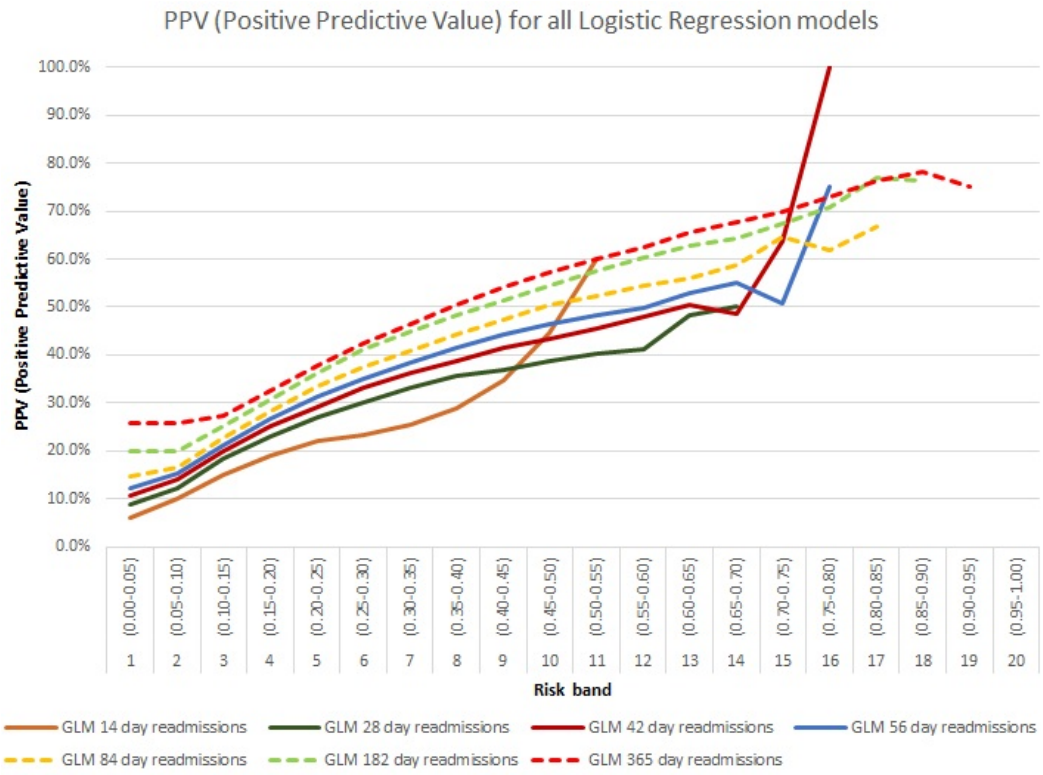

Figure 7.1: Positive Predictive Value of all Logistic Regression models

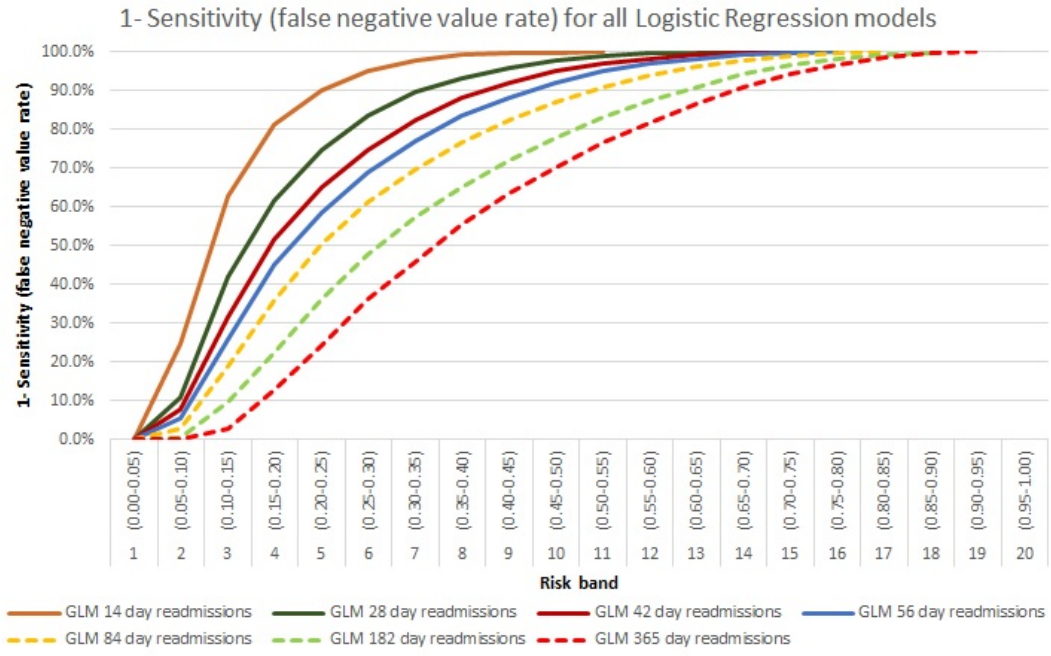

Figure 7.2: False Negative Rate of all Logistic Regression models

not predicting a lot of patients in them, they are predicted correctly. The 14 day readmissions model also has a higher NPV value than the other periods meaning low risk patients are predicted correctly as low risk.

For the remaining measures sensitivity and specificity the model did not perform as well as the longer time periods. The results reinforce the idea that if the time period is too short the rate of false negatives will be too high as seen in Figure 7.2. This means the proportion of patients that are incorrectly predicted as low risk is much larger than the proportion that are correctly predicted as at risk. Also note that this rate for all readmission band time periods eventually reaches 100%, this is because the number of the patients predicted at low risk is high.

7.2.2 28 days

The results of the 28 day readmission model are in Table (7.2). At a risk threshold of 0.5 the model only predicts 456 patients as at risk of readmission. That is, over 4 years and 4 months of data only approximately 1 patient every 3 days is at risk of readmitting to hospital. Consultation with DHB management about this model found this value is too small to be used within the DHB so for the 28 day readmission model the optimal risk threshold is at 0.4 predicting 2192 patients as at risk. A downfall of dropping to this threshold is the PPV reduces by approximately 3.3% as seen in Figure 7.1.

The performance measure results for this model show that although the DHB and MoH regard this period as the optimal period between initial discharge and readmission the PPV, sensitivity and specificity indicate otherwise. Compared to the PARR-30 model measures at a risk threshold of 0.5 this 28 day readmission model does not compare well. PARR-30 PPV was 59.2% compared this model, 40.1%. The sensitivity and specificity of the PARR-30 model was 5.4% and 99.5% whereas this model the sensitivity was very low

at 0.9% and the specificity was greater than PARR-30 at 99.9%. The ROC C-Statistic calculated during cross validation in this study was 0.74 which is slightly better than the PARR-30 model value of 0.7. Dropping the risk threshold to 0.4 increases the total patients at risk but it decreases the PPV which is not desirable.

7.2.3 42 days

The PPV for this model is greater than the two shorter time period models. The results of this time period indicate that this 6 week time period performs in between the 28 and 56 day models which is shown in the PPV, sensitivity and specificity figures.

Table 7.5: Logistic Regression Model risk band table 56 days

Risk Band	Probability	Total at risk	% of total	Total readmit	Total do not readmit	% actually readmit	Total cost of readmission	Average cost of readmission	Total 28 day readmissions
1	(0.00-0.05)	54116	23.67%	1520	52596	2.8%	$3646332	$2399	913
2	(0.05-0.10)	75676	33.11%	5692	69984	7.5%	$33172465	$5828	4151
3	(0.10-0.15)	41154	18%	5465	35689	13.3%	$37287865	$6823	3981
4	(0.15-0.20)	20550	8.99%	3774	16776	18.4%	$27472971	$7280	2656
5	(0.20-0.25)	12140	5.31%	2897	9243	23.9%	$20717072	$7151	2032
6	(0.25-0.30)	8138	3.56%	2261	5877	27.8%	$15659185	$6926	1609
7	(0.30-0.35)	5621	2.46%	1840	3781	32.7%	$13448111	$7309	1285
8	(0.35-0.40)	3794	1.66%	1374	2420	36.2%	$9820477	$7147	972
9	(0.40-0.45)	2595	1.14%	1040	1555	40.1%	$7594699	$7303	740
10	(0.45-0.50)	1876	0.82%	810	1066	43.2%	$5090887	$6285	583
11	(0.50-0.55)	1177	0.51%	541	636	46%	$3432083	$6344	393
12	(0.55-0.60)	838	0.37%	393	445	46.9%	$2515062	$6400	288
13	(0.60-0.65)	600	0.26%	309	291	51.5%	$2299489	$7442	232
14	(0.65-0.70)	258	0.11%	145	113	56.2%	$1319955	$9103	105
15	(0.70-0.75)	55	0.02%	27	28	49.1%	$347175	$12858	20
16	(0.75-0.80)	4	0%	3	1	75%	$32034	$10678	3
17	(0.80-0.85)								
18	(0.85-0.90)								
19	(0.90-0.95)								
20	(0.95-1.00)								
		228592	100%	28091	200501	12.3%	$183855862	$6545	

7.2.4 56 days

The risk band table for the 56 days between initial admission and readmission period can be found in Table 7.5. For this time period the optimal threshold is 0.5 as the PPV is high at this level, 48.36%, and the total predicted at risk is 2932 patients. The PPV is highest in this model compared to the other short time period models indicating that this 56 day period is optimal for predicting patients at risk of readmission.

The correctly identified at risk patients over the total at risk (the sensi-

tivity) is the highest in this readmission period model out of the four short time period models Figure 7.3. Alternatively the specificity which is the true negative rate is lowest in the 56 day model (Figure 7.4) compared to the 14, 28 and 42 day models.

The PPV for this 56 day period model at a risk threshold of 0.5 (48.36%) compared to the PARR-30 model (59.2%) was a lot better than the 28 day model, although not as high as we would have liked compared to PARR-30. Fortunately the sensitivity for this model,5.0%, and PARR-30, 5.4%, are nearly the same. The same pattern occurs with the specificity which is lower in the 56 model (99.2%) compared to the 28 model (99.9%). The ROC value did not change from the 28 model, 0.74, which indicates that the model is performing better than a random guess.

For this model the number of 28 day readmissions was also calculated to see whether the number predicted were comparable to the number predicted at risk in the 28 day model. At a risk threshold of 0.5 the total patients predicted at risk within 28 days is 1041 patients in the 56 day model. This is more than double the patients predicted in the 28 day model (456 at risk). This means the 56 day model is correctly predicting patients at risk within 28 days better than the 28 day model. This is because true positives in the 28 day model is 183 at the 0.5 threshold and 807 at the 0.4 threshold. Both of these values are lower than the 56 day model (1041) indicating that the 56 day model is optimal in terms of predicting the number of patients who readmit within 28 days.

7.2.5 84, 182 and 365 days

The number at risk and the PPV for the 12 weeks, 6 months and 1 year time periods between the initial admission and the readmission are displayed in Ta-

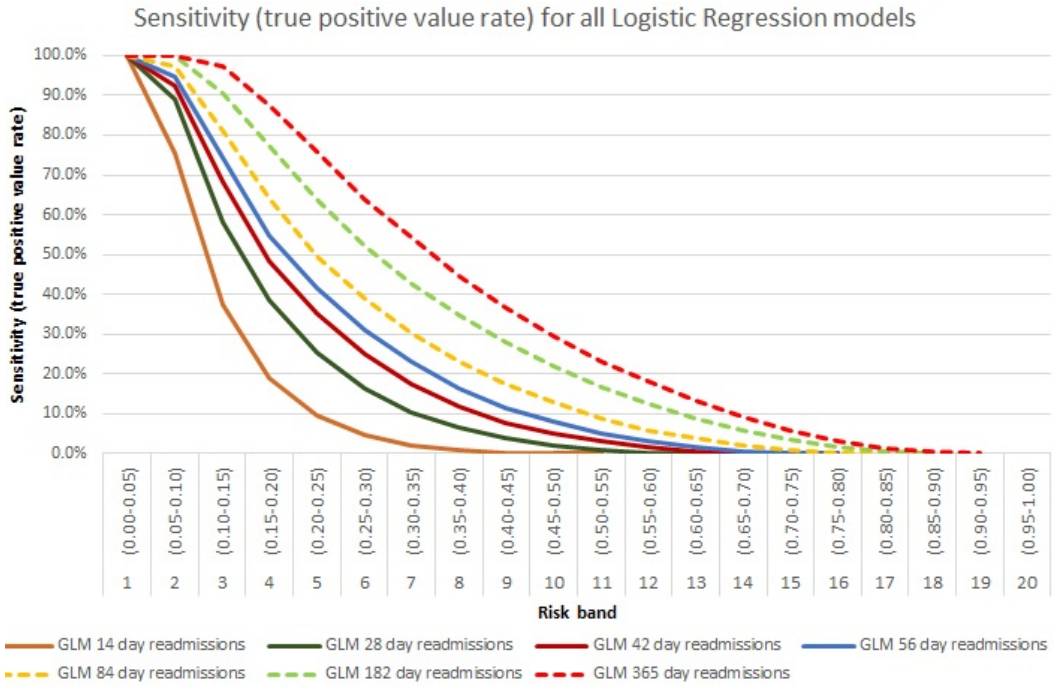

Figure 7.3: Sensitivity of all Logistic Regression models

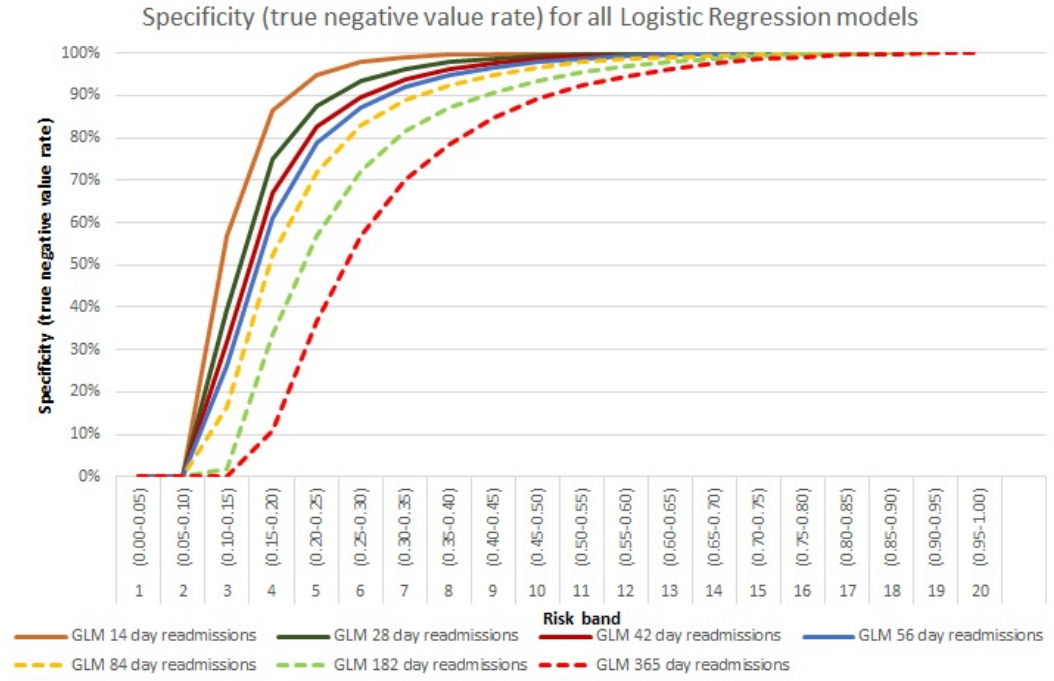

Figure 7.4: Specificity of all Logistic Regression models

ble 7.4. The number at risk for all of these models indicate a greater number of patients at risk as the readmission period is increased. This is because as the time between the two instances rises more patients are likely to readmit but it is less likely that the readmissions that are flagged by the Waikato DHB data are related to the initial admission. This is because the data defined readmission may actually be ongoing medical concerns or accidents unrelated to their initial admission. It is important to keep that in mind when looking at the PPV for these long readmission period models compared to the shorter time periods. Although it may look better, it may be masking the false positives (actual readmissions that are unrelated to the initial admission). Eventually, if the time period between initial admission and readmission is long enough, all of the patients in the initial dataset are highly likely to readmit. This is evident if you compare the total that readmit across risk bands over the different time periods as the centre of the distribution of total that readmit shifts towards the higher risk bands.

The the sensitivity (Figure 7.3) and PPV (Figure 7.1) are highest in these models compared to the shorter time period models. Alternatively the specificity which is the true negative rate is lowest for these models 7.4) compared to the 14, 28, 42 and 56 day models.

Although 12 month period is not the model of interest in this thesis it is interesting to compare the PPV to the results of the Vaithianathan et al. (2012) study. Using a risk score threshold of 0.7 for this model resulted in a PPV of 73.37% and at a threshold of 0.9 the PPV was 91.67%. Compared to this study the PPV was reasonably similar for a risk threshold of 0.7 68.7% and for 0.9 75.0% however very few were predicted in this high cut off band.

7.2.6 Risk Threshold and Readmission Days Summary

Waikato DHB management came to the conclusion that the 56 day readmission period and risk threshold of 0.5 are the optimal predictive criteria for predicting patients at risk of a readmission. This is clearly demonstrated in the PPV graph of all time periods (Figure 7.1). The 56 day readmission model also correctly predicts patients who readmit within 28 days better than the 28 day model as described in Section 7.2.4. The results of this analysis also show that the PPV is the most important criteria as it is a measure which describes the number of correctly identified high risk patients readmitting over the total at risk in each risk band. This measure has more practical value to the DHB and has more meaning the sensitivity which is the proportion of patients at risk that readmit over the total readmissions. This is because it is more desirable for the model to be predicting accurately than predicting many at risk.

7.3 Naive Bayes

The Naive Bayes model was fit to the data using the naive.bayes function from the bnlearn package in R. This algorithm uses the readmission as the class variable and calculated the posterior probability using Bayesian estimation for each of the observations in the dataset. The risk band tables were created for the Naive Bayes models for four different initial discharge to readmission time periods; 14, 28, 42 and 56 days. A summary table of the number at risk and PPV for Naive Bayes is found in Table 7.6.

The differences in the performance measures between the four different readmission time periods perform similarly to the Logistic Regression models. For example in Figure (7.5) the PPV is lowest for the short 14 model and largest in the 56 day model. The sensitivity in Figure (7.6) is highest again in the 56 model and lowest in the 14 day model.

Naive Bayes does not perform as well as the Logistic regression models, although it is predicting a high number of patients in the high risk bands (Table 7.6). This means that the sensitivity is greater amongst the Naive Bayes models compared to the Logistic regression models, seen in Figure 7.6. This is because the number of true positives is greater in the Naive Bayes models. This is attributed to the model predicting more patients at risk (there are more patients in the high risk bands) so the percentage of true positives over the total actual readmissions is higher. But these patients are not correctly identified, therefore the Naive Bayes models all have a low PPV (Figure 7.5). For the 56 day Naive Bayes model the PPV is only 30.8% at the 0.5 threshold and still only 36.3% at the 0.75 threshold. These PPVs are low compared to the logistic regression model for the same period which are 48.4% for 0.5 threshold and 75.0% for 0.75 threshold. This pattern is seen throughout all of the Naive Bayes models over the different time periods. This highlights the problem with using the sensitivity and, similarly, the specificity as performance measures as they mask the true predictive power of a model. Performing an intervention on the patients that the Naive Bayes model predicts as at risk would mean the DHB is at a loss as they would be spending money on patients that are not actually likely to readmit.

Table 7.6: Naive Bayes Model total at risk and PPV by risk threshold and 14, 28, 42 and 56 readmission days

Risk	Probability	14 days		28days		42 days		56 days	
Band		At risk	PPV	At risk	PPV	At risk	PPV	At risk	PPV
1	(0.00-0.05)	233334	5.9%	233334	8.7%	228592	10.7%	228592	12.3%
2	(0.05-0.10)	82795	11%	97524	15.3%	103068	18.1%	107671	20.2%
3	(0.10-0.15)	59483	12.4%	73694	17.2%	78119	20.6%	82662	22.9%
4	(0.15-0.20)	48018	13.3%	59451	18.7%	65048	22.1%	70450	24.5%
5	(0.20-0.25)	39308	14.1%	51253	19.6%	56065	23.4%	60456	26.1%
6	(0.25-0.30)	33624	14.8%	44527	20.6%	49700	24.4%	53399	27.2%
7	(0.30-0.35)	28986	15.4%	38995	21.4%	44441	25.3%	48157	28.2%
8	(0.35-0.40)	25100	15.9%	34816	22.1%	39222	26.2%	43280	29.2%
9	(0.40-0.45)	22040	16.3%	31007	22.8%	35592	26.9%	38912	30%
10	(0.45-0.50)	19196	16.7%	27536	23.7%	31972	27.8%	35447	30.8%
11	(0.50-0.55)	16618	17%	24550	24.3%	28904	28.5%	31935	31.7%
12	(0.55-0.60)	14355	17.3%	21897	24.8%	25614	29.6%	28865	32.6%
13	(0.60-0.65)	12461	17.6%	19428	25.5%	23069	30.2%	25833	33.4%
14	(0.65-0.70)	10518	17.9%	16868	26.2%	20443	31%	23120	34.3%
15	(0.70-0.75)	8656	18%	14449	26.9%	17752	31.8%	20401	35.2%
16	(0.75-0.80)	7023	18.3%	12335	27.5%	15081	32.8%	17581	36.3%
17	(0.80-0.85)	5358	18.8%	10003	28.3%	12661	33.7%	14667	37.6%
18	(0.85-0.90)	3715	19.6%	7602	29%	9963	34.6%	11877	38.9%
19	(0.90-0.95)	2237	21.5%	5191	29.9%	7067	36%	8592	39.9%
20	(0.95-1.00)	891	23.6%	2598	33%	3746	38.1%	4828	42.3%

7.4 Cost Analysis

The cost analysis technique described in Section 6.4 was performed on both the logistic regression and Naive Bayes models over the readmission periods 14, 28, 42 and 56 days. The average cost of the readmission seen in the 28

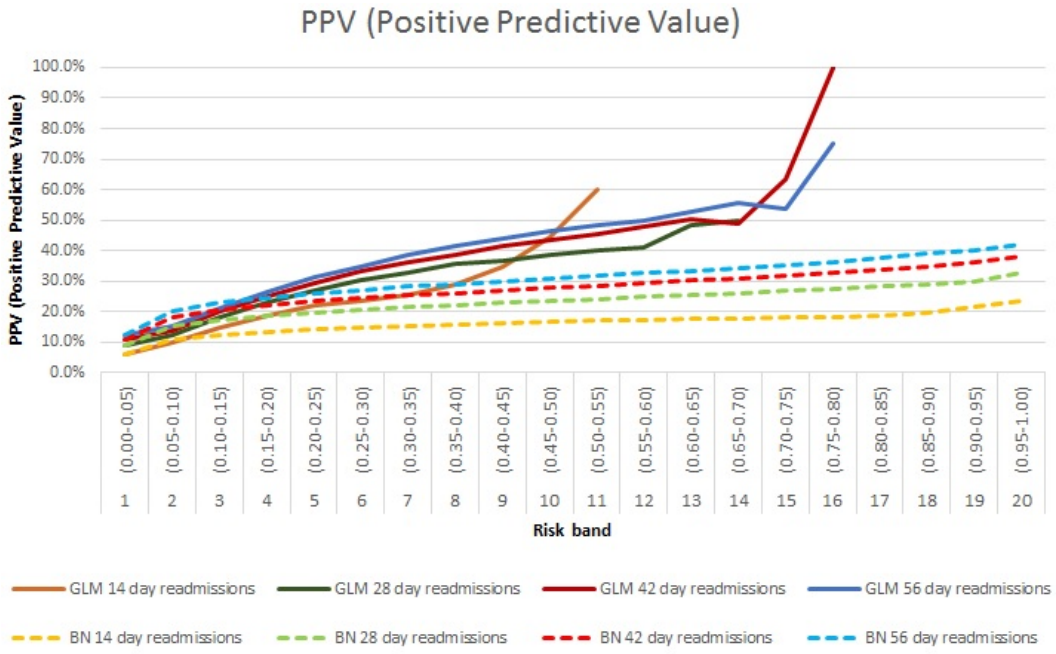

Figure 7.5: Positive Predictive Value for Logistic Regression and Naive Bayes models

day (Table 7.2) and 56 day (Table 7.5) risk band tables is the average cost for a readmission for the patients that readmit only. It is evident that the mean readmission cost increases as the risk band probability increases. Figure (7.7) displays that cost for the logistic regression and Naive Bayes models. The cost increases rapidly at the 0.5 threshold mark for the logistic regression models which indicates that the patients that are the most expensive to the DHB are in the high risk bands. This is a positive result as the more resource intensive patients are predicted as high risk. The Naive Bayes models do not have the peak in cost as the logistic regression models do. This suggests that they are not predicting the expensive patients in the high risk bands as well as the logistic regression models are. This again confirms that the logistic regression models perform better than the Naive Bayes models.

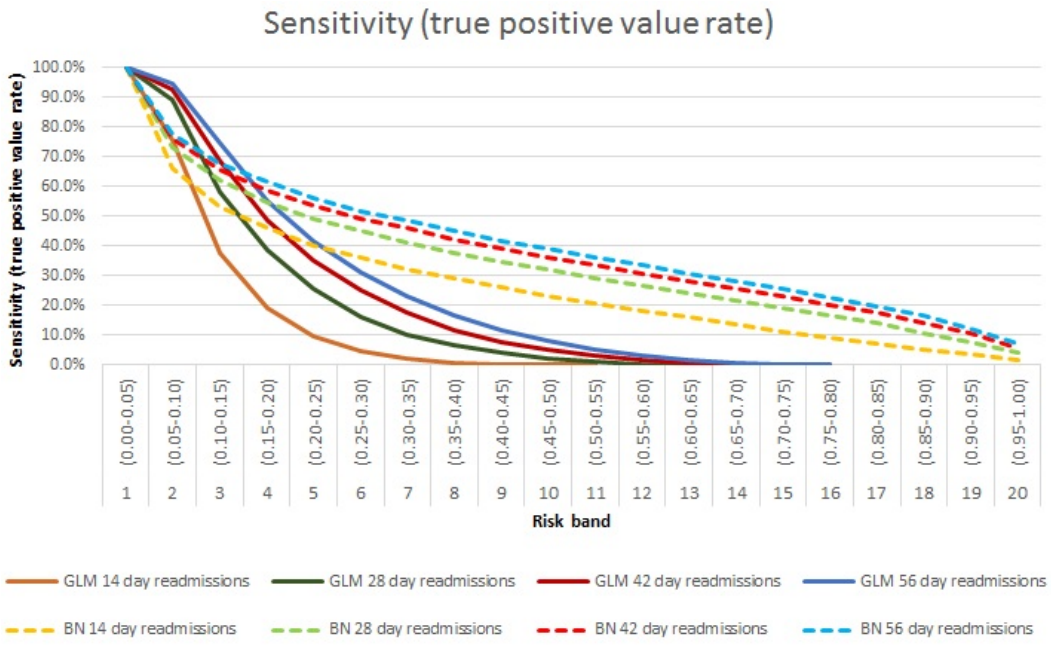

Figure 7.6: Sensitivity for Logistic Regression and Naive Bayes models

7.4.1 100% Readmission Reduction Cost Analysis

The three different intervention costs, $500, $1000 and $2000 were compared over the 20 different risk thresholds and the four readmission time periods 14, 28, 42 and 56 days for the logistic regression models only.

The $2000 intervention savings were not distinguishable between the different time periods although the savings did increase as the risk threshold increased. This suggests that an intervention cost of $2000 would be too high as it increases the cost of using the model without saving more money than the other low cost interventions. At a low risk threshold the model saves the most money but it would mean the model identifies next to everyone as at risk therefore the PPV is very low (Figure 7.10) . This would not be practical as the DHB would not want to spend the cost of an intervention on everyone who is discharged from hospital. A more practical risk threshold level would be 0.5 which although according to the analysis it does not save the most money.But the savings for the $500 and $1000 intervention costs are similar from a risk threshold of about 0.5 onwards as seen in Figures 7.8 and 7.9. Those figures also show that the 56 day readmission has the highest net savings for the DHB.

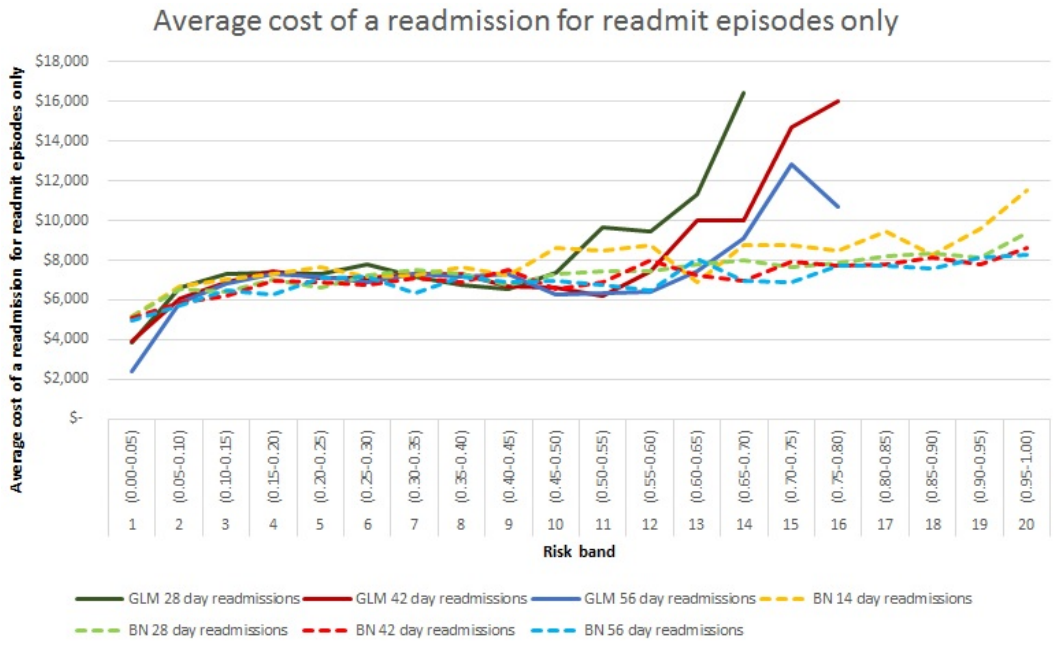

Figure 7.7: Average cost of a readmission for readmissions only

Paired with the PPV in figure 7.10 using an intervention of $1000 shows that since the savings are flattening out then PPV is at a good point at 0.5 as well.

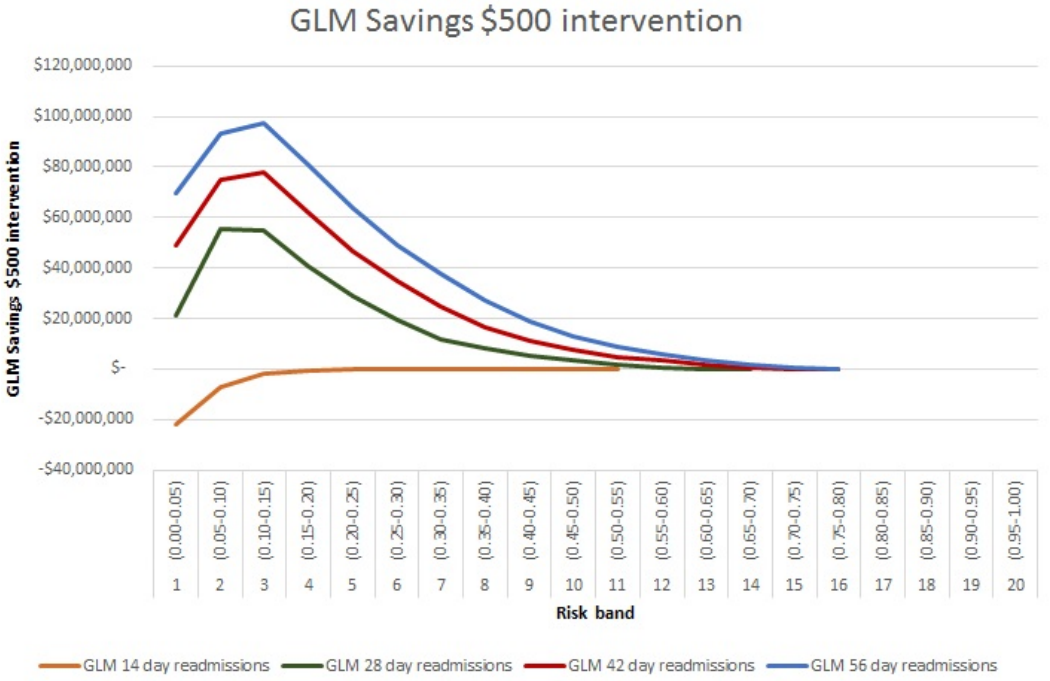

Figure 7.8: Cost savings for Logistic Regression models ($500 intervention)

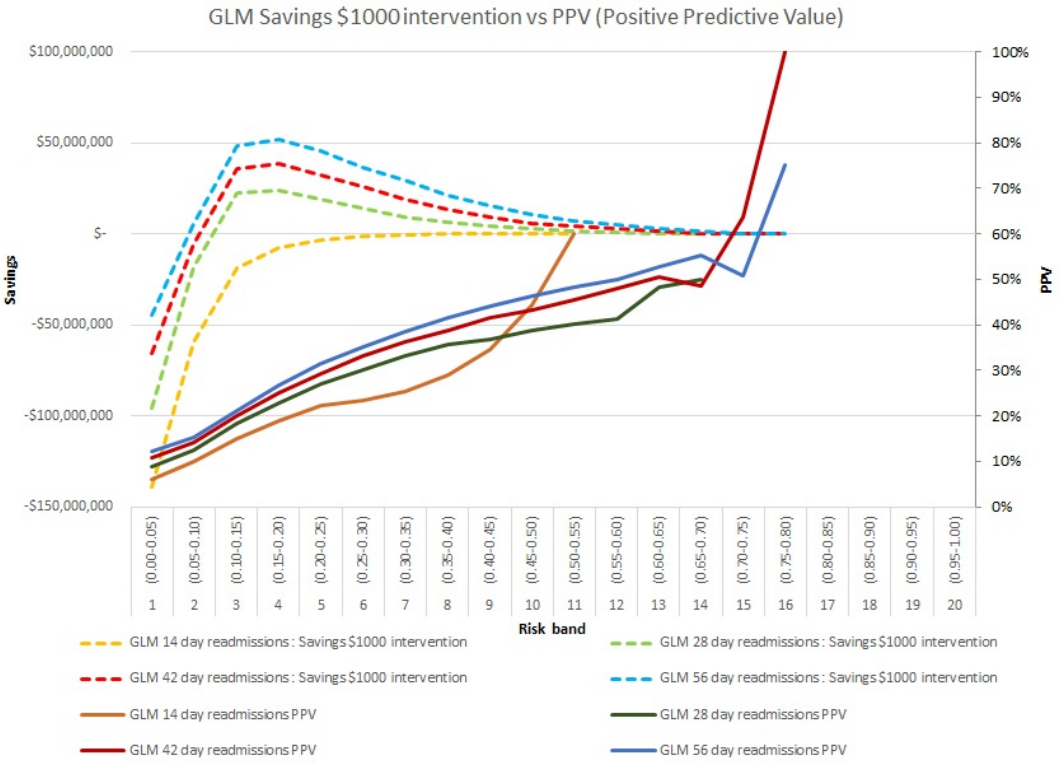

Figure 7.10: Cost savings for Logistic Regression models ($1000 intervention) versus PPV

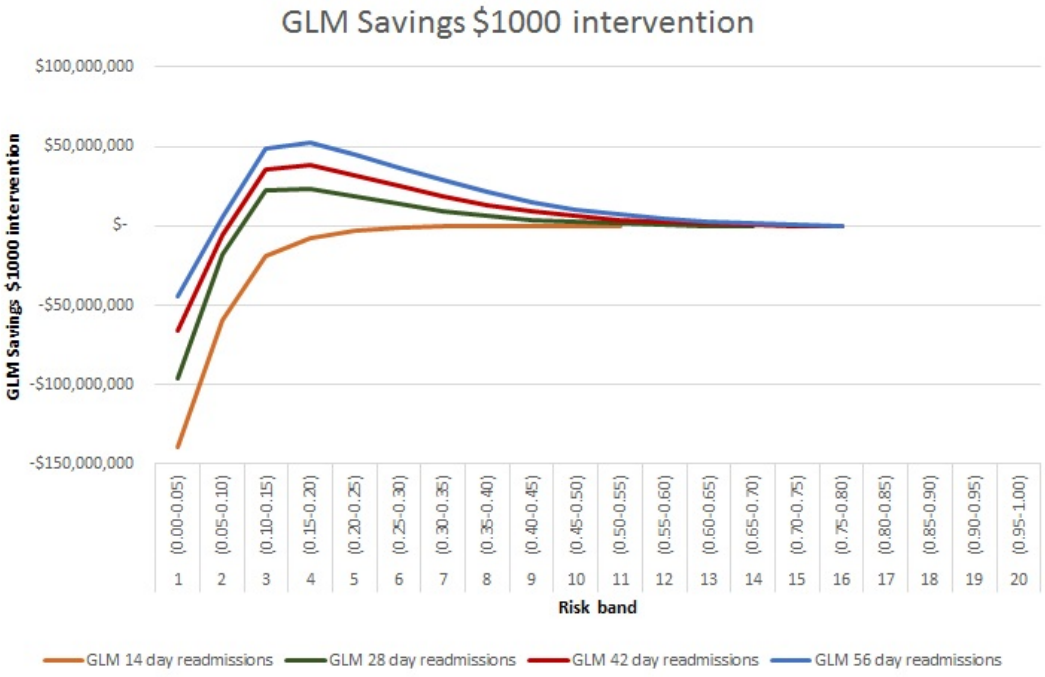

Figure 7.9: Cost savings for Logistic Regression models ($1000 intervention)

Band	Probability	$500 intervention			$1000 intervention			$2000 intervention		
		Cost savings for 10% reduction of true positives	Cost savings for 20% reduction of true positives	Cost savings for 50% reduction of true positives	Cost savings for 10% reduction of true positives	Cost savings for 20% reduction of true positives	Cost savings for 50% reduction of true positives	Cost savings for 10% reduction of true positives	Cost savings for 20% reduction of true positives	Cost savings for 50% reduction of true positives
1	(0.00-0.05)	-$ 95,910,414	-$ 77,524,828	-$ 22,368,069	-$ 210,206,414	-$ 191,820,828	-$ 136,664,069	-$ 438,798,414	-$ 420,412,828	-$ 365,256,069
2	(0.05-0.10)	-$ 69,217,047	-$ 51,196,094	$ 2,866,765	-$ 156,455,047	-$ 138,434,094	-$ 84,371,235	-$ 330,931,047	-$ 312,910,094	-$ 258,847,235
3	(0.10-0.15)	-$ 34,696,294	-$ 19,992,587	$ 24,118,533	-$ 84,096,294	-$ 69,392,587	-$ 25,281,468	-$ 182,896,294	-$ 168,192,587	-$ 124,081,468
4	(0.15-0.20)	-$ 17,848,080	-$ 6,873,160	$ 26,051,600	-$ 46,671,080	-$ 35,696,160	-$ 2,771,400	-$ 104,317,080	-$ 93,342,160	-$ 60,417,400
5	(0.20-0.25)	-$ 10,320,377	-$ 2,092,754	$ 22,590,115	-$ 28,868,377	-$ 20,640,754	$ 4,042,115	-$ 65,964,377	-$ 57,736,754	-$ 33,053,886
6	(0.25-0.30)	-$ 6,322,084	-$ 166,169	$ 18,301,579	-$ 18,800,084	-$ 12,644,169	$ 5,823,579	-$ 43,756,084	-$ 37,600,169	-$ 19,132,422
7	(0.30-0.35)	-$ 3,819,003	$ 770,994	$ 14,540,986	-$ 12,228,003	-$ 7,638,006	$ 6,131,986	-$ 29,046,003	-$ 24,456,006	-$ 10,686,014
8	(0.35-0.40)	-$ 2,353,314	$ 891,872	$ 10,627,431	-$ 7,951,814	-$ 4,706,628	$ 5,028,931	-$ 19,148,814	-$ 15,903,628	-$ 6,168,070
9	(0.40-0.45)	-$ 1,438,362	$ 824,777	$ 7,614,192	-$ 5,139,862	-$ 2,876,723	$ 3,912,692	-$ 12,542,862	-$ 10,279,723	-$ 3,490,308
10	(0.45-0.50)	-$ 900,332	$ 603,337	$ 5,114,343	-$ 3,304,332	-$ 1,800,663	$ 2,710,343	-$ 8,112,332	-$ 6,608,663	-$ 2,097,658
11	(0.50-0.55)	-$ 471,420	$ 523,160	$ 3,506,899	-$ 1,937,420	-$ 942,840	$ 2,040,899	-$ 4,869,420	-$ 3,874,840	-$ 891,101
12	(0.55-0.60)	-$ 226,129	$ 425,243	$ 2,379,358	-$ 1,103,629	-$ 452,257	$ 1,501,858	-$ 2,858,629	-$ 2,207,257	-$ 253,143
13	(0.60-0.65)	-$ 58,635	$ 341,231	$ 1,540,827	-$ 517,135	-$ 117,269	$ 1,082,327	-$ 1,434,135	-$ 1,034,269	$ 165,327
14	(0.65-0.70)	$ 11,416	$ 181,333	$ 691,082	-$ 147,084	$ 22,833	$ 532,582	-$ 464,084	-$ 294,167	$ 215,582
15	(0.70-0.75)	$ 8,421	$ 46,342	$ 160,105	-$ 21,079	$ 16,842	$ 130,605	-$ 80,079	-$ 42,158	$ 71,605
16	(0.75-0.80)	$ 1,203	$ 4,407	$ 14,017	-$ 797	$ 2,407	$ 12,017	-$ 4,797	-$ 1,593	$ 8,017
17	(0.80-0.85)									
18	(0.85-0.90)									
19	(0.90-0.95)									
20	(0.95-1.00)									

Figure 7.11: 56 Day Readmission Model Cost Analysis for 10%, 20% and 50% Readmission Reductions

7.4.2 10%, 20% and 50% Readmission Reduction Cost Analysis

The previous cost analysis is based on the assumption that the model has perfect efficacy as it is assumed that no patients readmit. The second form of cost analysis was performed on the 56 day model only. The assumption that 100% of patients identified as at risk not returning is unlikely so in this section different readmission reduction rates are tested. The pessimistic reduction to only 10% would mean 10% of the patients that the model identifies as high risk do not return, meaning the other 90% of flagged patients do have a readmission. Also used was a 20% reduction in readmissions and a more optimistic reduction of 50%. This means the true positives are multiplied by the reduced rates and the savings from the model are recalculated at each of the three reduction levels and the three intervention costs. The results for the 56 day readmission model can be found in the table in Figure 7.11, the cost savings in green indicate that the model will be making a saving if used by the DHB based on the assumptions of the analysis. This shows that if only 10% of patients the model predicts as at risk do not readmit then the DHB would not save money no matter how little they spend on intervention. The only net savings for the DHB would be using a $500 intervention and that would result in a saving of $11,416 using a risk threshold of 0.65 which would not be beneficial for the DHB as this sum minimal. For the optimistic reduction level of 50% the DHB would save approximately $3.5 million for an intervention cost of $500. For a reduction of 20% the savings would be approximately $500,000 for the DHB using $500 per patient for intervention. The results in Table (7.11) do indicate that the $500 intervention cost would be the optimal cost for the DHB for an intervention cost per patient. If the reduction in readmissions is lower than the overly optimistic 100% we assumed previously then the intervention cost of $500 saves the DHB the most money.

7.5 Conclusions

The models, statistical and predictive criteria and cost analysis techniques described in this chapter were used to find the optimal model for predicting patients at risk of readmission. The statistical results of this analysis indicate that the logistic regression model using the explanatory variables summarised in Section 7.1.1 fit the readmission data the best. Given this model the time period and risk threshold with the optimal criteria for predicting patients at risk of readmission is the 56 day model at a risk threshold cut off of 0.5. This conclusion is reached by focusing on both the cost analysis and the PPV results (Figure 7.10). This model identifies approximately 3,000 patients at risk which equates to about 2 patients at risk per day in the dataset described. The cost savings at this day period, intervention cost and risk threshold would be approximately $7 million if an intervention cost of $1000 was used.

Chapter 8

Discussion and conclusions

Readmissions are acute, unplanned admissions to hospital within a defined period of time from an initial admission. Readmission rates are a well established health quality measure both in New Zealand and internationally as some readmissions that do occur are avoidable and if data is modelled correctly groups of patients at high risk of readmission are identifiable. The MoH in New Zealand and other international health governing bodies use 28-30 days between the initial admission and the acute readmission as their optimal readmission time period. This is because it is considered the most likely time frame that the two events are related. A time frame of approximately 12 months would increase the chances of picking up admissions unrelated to the initial admission.

Case finding algorithms such as predictive risk modelling attempt to identify patients at high risk of readmission. Research indicates predictive risk models have been used for some time and have developed significantly. Studies internationally such as the PARR-30 models in the UK develop an algorithm to identify patients at high risk of readmitting to hospital within 30 days of discharge. Majority of these studies use logistic regression models to build these predictive risk models.

This thesis is concerned with developing a predictive risk model to identify

patients that are at high risk of readmission to hospital using Waikato DHB data. A range of different criteria is considered to find the optimal model. We wish to identify the optimal time period between initial admission and readmission and the optimal risk threshold for predictive modelling purposes. To our knowledge no studies address whether the period of 28 days between initial discharge and readmission is optimal for predictive risk modelling.

8.1 Discussion

Logistic regression models were fit to the 28 day readmission data. The model that fits the data best was calculated using the model AIC and residual deviance and the cross validation performance measures (the PPV, sensitivity, specificity, ROC C-Statistic and the total at risk). The best model in terms of variable selection and model fit includes these 21 variables: fiscal year, discharge specialty cluster group, LOS group, age group, patient category, admit type, 13 diseases (Myocardial infarction, Congestive heart failure, Peripheral vascular disease, Chronic pulmonary disease, Mild liver disease, Diabetes, Moderate or severe renal disease, Diabetes with end-organ damage, Leukaemia, Lymphoma, Moderate or severe liver disease, Metastatic solid tumour and AIDS/HIV), the number of acute admissions in the previous 12 months grouped and the number of ED presentations in the previous 12 months from the initial admission grouped. All variables were of clinical significance.

The results of this 28 day readmission risk model in comparison to the PARR-30 model discussed previously indicate that this 28 day model does not perform as well as PARR-30. The PPV, 40.1%, is significantly lower than the PARR-30 model, 59.2%. The specificity of 99.9% is only slightly greater than PARR-30 99.5% and the sensitivity in this model is very low, 0.9%, compared to PARR-30, 5.4%. The ROC C-Statistic, 0.74, is similar to PARR-30, 0.70.

Using the explanatory variables identified as significant the model is fit to data over a range of readmission periods: 14, 28, 42, 56, 84, 182 and 365 days. This is done to test whether the 28 day period is in fact the optimal time period between initial discharge and readmission as the literature and reporting measures assume. The risk threshold of 0.5, widely used in the literature, is also assessed to see if this is the optimal cut off point to identify patients at risk of readmission.

Results show that the 14 day period is too short as it does not predict many patients in the high risk bands. At the opposite end of the scale the 365 day model predicts the most patients in the high risk bands and has a high PPV value. This is because as time between the initial discharge and readmission increases, more patients are likely to readmit as over a long enough period of time all patients may readmit. But it is less likely that these readmissions are acute readmissions related to the initial admission. This is because the data defined readmission may actually be ongoing medical concerns or accidents unrelated to their initial admission. That is important to consider when looking at the number at risk and PPV for these long readmission period models compared to the shorter time periods and it is for that reason 84 and more days are not considered further in this thesis.

The model fit to the 56 day data has an optimal risk threshold at 0.5 as the PPV is highest at this level, 48.36%, and the total predicted patients at risk, 2932, is a manageable number of patients for intervention. The PPV is highest in this model compared to the other short time period models (14, 28 and 42 days) which indicates that this 56 day period is optimal for predicting patients at risk of readmission. The PPV for this 56 day period model at a risk threshold of 0.5 is low (48.36%) compared to the PARR-30 model (59.2%) but it is better than the 28 day model PPV at the same threshold. The sensitivity

for this model, 5.0%, is very similar to the PARR-30 result of 5.4%.

Analysis of the 56 day model also found that it is correctly predicting patients who readmit within 28 days better than the 28 day model itself. This is because the true positives in the 28 day model is 183 at the 0.5 threshold and 807 at the 0.4 threshold. Both of these values are lower the number that the 56 day model predicts for patients who readmit within 28 days, 1041. This indicates that, based on the threshold and performance measures described previously, the 56 day model is optimal in terms of predicting the number of patients who readmit within both 28 days and 56 days.

Naive Bayes does not perform as well as the Logistic regression models. The models are predicting a large number of patients in the high risk bands but these patients are not being correctly identified. This results in a low PPV for all of the Naive Bayes models. For the 56 day Naive Bayes model the PPV is only 30.8% at the 0.5 threshold and still only 36.3% at the 0.75 threshold. These PPVs are low compared to the logistic regression model for the same period which are 48.4% for 0.5 threshold and 75.0% for 0.75 threshold. This pattern of results is seen throughout all of the Naive Bayes models over the different time periods. Using the Naive Bayes model for risk prediction would mean the DHB would be spending money on patients that are not actually likely to readmit.

The cost analysis in this thesis reinforced the use of the 56 day model for risk prediction. This model has the highest net savings for the DHB at all intervention costs if the assumption that all patients that are predicted to readmit do not readmit is correct. A second cost analysis was performed on the 56 day model to see what the effect on the cost savings is and what the optimal risk threshold would be if only 10%, 20% and 50% of patients identified as at risk do not readmit (compared to the first analysis where 100% are assumed

to not readmit). The results indicate that the $500 intervention cost would be the optimal cost for the DHB to spend on an intervention per patient. For the optimistic reduction level of 50% the DHB would save approximately $3.5 million and if patients that do not readmit reduced to 20% the savings would be approximately $500,000 for the DHB. Note that this is over a period of approximately 4 years and 4 months. The cost analysis performed on the models indicate the low intervention cost of $500 is the most cost effective expenditure. This is evident when comparing both the savings when the model assumes 100% of the patients identified as at risk do not readmit as well as only 10%, 20% and 50% not readmitting.

8.2 Recommendations

8.2.1 How to Measure a Readmission

A very important recommendation from this study for the Waikato DHB is to define readmissions in the data correctly. At the moment readmissions are defined by a set of data qualifications run on a large dataset, as it is at the MoH and most likely many other health organisations. Therefore the readmission we use in this analysis is a data qualified readmission, not necessarily a true readmission that is related to a previous admission. The problem is there is no way of knowing whether the two events are related. It is for that reason it is recommended that the DHB introduce a method where hospital staff who admit patients are to flag whether an episode for a patient is a readmission related to a previous admission or not. This would decrease the uncertainty in the data and could potentially reduce the actual readmission numbers as the data qualifications may be exaggerating readmission rates. This would effect the predictive model significantly as the actual readmission rate is likely to be reasonably smaller than the current data qualified readmission rate. This could mean the predictive risk model developed in this study could in fact be predicting patients who are true readmissions. This could be why we have a

PPV sitting at approximately 50%. Correctly identifying readmissions could mean an increase in our PPV and other performance measures. It is also important to bear in mind that the other predictive risk studies discussed in this thesis also face the same issues.

8.2.2 Investigate the Efficacy of Intervention

Another recommendation is that the DHB test this model over a short trial period to track if patients that have an intervention (because the model predicts them as high risk). This could be monitored on a daily or weekly basis. This is because it is important to see if the intervention is effective and if so, to what extent. This is the proportional of patients identified by the model as high risk that do not readmit. This would verify the readmission reduction rates that are assumed in the cost analysis in Section 7.4.2. This may help verify the type of intervention required and the total cost that can be spent on that intervention per patient. This could be either low cost, such as monthly visits to a GP in primary care over a year (approximately $500), or a longer stay in hospital (approximately one day at $1000 per day).

8.2.3 Cost Analysis

A recommendation would also be to delve more into the cost analysis in this study. This is highlighted in the results section where different readmission reduction rates are assumed based on the likelihood that the DHB is unlikely to achieve a 100% reduction on the total the readmit. More research and business case analysis would have to be done to justify the resource needed for interventions on the identified high risk patients.

8.2.4 Social Factors

Social factors were not included in this study due to lack of information. Studies such as Hasan et al. (2010) were able to utilise information about the social support a patient did or did not have. Social support variables are likely to effect whether a patient readmits to hospital or not. Particularly for elderly patients that do not have a spouse or rest home level care after a stay in hospital as they are likely to be reliant on others for daily tasks. Unfortunately this information is not readily available in New Zealand as hospital information systems only collect data on a patients stay in hospital and information regarding that stay. However, there is potential for further analysis as in New Zealand we have the NHI, a unique identifier which has the potential to link many different types of information back to a singular identification number. It may be feasible to link the information in our model to the rest home data that the Waikato DHB collects. This is a recommendation for the future as the model could specifically look at an older age group and predict their risk considering additional social factors such as whether they reside in a rest home and if so, what level of care is available to them. It would also be beneficial if we could link other information such as Primary care data, Corrections, Education and Social Development data through the NHI number back to DHB level information. But this is dependent on confidentiality and data ownership issues.

This thesis found that the ethnicity and deprivation score explanatory variables were not significant. This is surprising considering those types of variables are renowned in New Zealand as variables that do impact hospitalisation. This indicates more investigation needs to be done around creating explanatory variables that incorporate social and cultural effects of hospital admission.

8.3 Concluding Remarks

The statistical and predictive criteria and cost analysis techniques described in this thesis were used to find the optimal model for predicting patients at risk of readmission. The statistical results of this analysis indicate that the logistic regression model using the explanatory variables summarised in Section 7.1.1 fit the 28 day readmission data the best. The logistic regression model correctly predicts high risk patients better than the Naive Bayes model, therefore it is the preferable classification technique.

Using the logistic regression model the optimal time period and risk threshold is found for predicting patients at risk of readmission. 28 days is no the optimal criteria based on the results of this study as the MoH and other health organisations believe. This study indicates that the 56 day readmission model with a risk threshold of 0.5 is the best. This conclusion is reached by focusing on both the cost analysis and the PPV results (Figure 7.10). This model has the greatest cost savings (if we assume a 100% reduction in readmissions) and also the highest PPV indicating it predicts patients at risk better than the other models discussed in this thesis.

The 0.5 cut off point identifies approximately 3,000 patients at risk which equates to about 2 patients per day for intervention over the 4 years and 3 months period that is modelled. Assuming 100% of the patients identified as at risk do not return the cost savings using the 56 day readmission model with an intervention cost of $500 per patient and risk threshold of 0.5 would be approximately $8.5 million. If the patients identified as a risk that do not return reduced to 50% not returning the DHB would save approximately $3.5 million for an intervention cost of $500. For a reduction of 20% the savings would be approximately $500,000 for the DHB for a $500 intervention per patient.

An important recommendation from this study is to find an objective way

to link readmissions to the initial episode. This would involve physically flagging patients as readmissions related to previous admission rather than relying on data qualified readmissions. Another important recommendation out of this thesis is that the efficacy of the intervention is investigated to verify what intervention works the best.

References

Australian Institute of Health and Welfare (2014). Hospital performance indicators.

Barber, D. (2012). *Bayesian reasoning and machine learning* (First ed.). univeristy Printing House, Cambridge CB2 8BS, United Kingdom: Cambirdge University Press.

Billings, J., Blunt, I., Steventon, A., Georghiou, T., Lewis, G., & Bardsley, M. (2012). Development of a predictive model to identify inpatients at risk of re-admission within 30days of discharge (PARR-30). *BMJ Open*, *2*(4).

Billings, J., Dixon, J., Mijanovich, T., & Wennberg, D. (2006). Case finding for patients at risk of readmission to hospital: development of algorithm to identify high risk patients. *BMJ*, *333*(7563).

Bottle, A., Aylin, P., & Majeed, A. (2006). Identifying patients at high risk of emergency hospital admission: a logistic regression analysis. *Journal of the Royal Society of Medicine*, *99*, 406–414.

Choudhry, S., Li, J., Davis, D., Erdmann, C., Sikka, R., & Sutariya, B. (2013). A public-private partnership develops and externally validates a 30-day hospital readmission risk prediction model. *Online Journal of Public Health Informatics*, *5*(2).

Dobson, A. & Barnett, A. (2008). *An Intorduction to Generalized Linear Models* (Third ed.). Boca Raton, FL, USA: Taylor & Francis Group.

Drozda, J. (2013). Readmission rates: edging slowly towards a deeper understanding and ultimately better care for patients. *BMJ*, *347*.

Gabbe, B., Harrison, J., Lyons, R., Edwards, E., & Cameron, P. (2013). Comparison of measures of comorbidity for predicting disability 12-months post-injury. *BMC Health Services Research*, *13*(1), 30.

Hasan, O., Meltzer, D., Shaykevich, S., Bell, C., Kaboli, P., Auerbach, A., Wetterneck, T., Arora, V., Zhang, J., & Schnipper, J. (2010). Hospital readmission in general medicine patients: a prediction model. *25*(3),

211–219.

Hersh, A., Masoudi, F., & Allen, L. (2013). Postdischarge environment following heart failure hospitalization: Expanding the view of hospital readmission. *Journal of the American Heart Association*, *2*(2).

New Zealand Ministry of Health (2012). DHB non-financial monitoring framework and performance measures.

Nguefack-Tsague, G. (2011). Using bayesian networks to model hierarchical relationships in epidemiological studies. *Epidemiology and Health*, *33*.

NHS National Services Scotland (2011). A report on the development of sparra (scottish patients at risk of readmission and admission).

Perlis, R. (2013). A clinical risk stratication tool for predicting treatment resistance in major depressive disorder. *Biological Psychiatry*, *74*, 7–14.

Robinson, T. & Kerse, N. (2012). Medical readmissions amongst older New zealanders: a descriptive analysis. *NZ Medical Journal*, *125*, 24–34.

Rumball-Smith, J., Sarfati, D., Hider, P., & Blakely, T. (2013). Ethnic disparities in the quality of hospital care in new zealand, as measured by 30-day rate of unplanned readmission/death. *International Journal for Quality in Health Care*, *25*(3), 248–254.

Rumball-Smith, J. & Hider, P. (2009). The validity of readmission rate as a marker of the quality of hospital care, and a recommendation for its definition. *NZ Medical Journal*, *122*, 6370.

Sarfati, D., Tan, L., Blakely, T., & Pearce, N. (2011). Comorbidity among patients with colon cancer in new zealand. *The New Zealand Medical Journal*, *124*(1338), 76–88.

Vaithianathan, R., Jiang, N., & Ashton, T. (2012). A model for predicting readmission risk in new zealand.

Witten, I. & Frank, E. (2005). *Data Mining: Practical Machine Learning Tool and Techniques* (Second ed.). San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.