A novel framework for predicting patients at risk of readmission

Rathi, M.

A NOVEL FRAMEWORK FOR PREDICTING PATIENTS AT RISK OF READMISSION

MANISHA RATHI

A thesis submitted in the partial fullfillment for the requirements for the degree of Doctor of Philosophy

DECEMBER 2, 2015

UNIVERSITY OF WESTMINSTER
Faculty of science and technology

Abstract

Uncertainty in decision-making for patients' risk of re-admission arises due to non-uniform data and lack of knowledge in health system variables. The knowledge of the impact of risk factors will provide clinicians better decision-making and in reducing the number of patients admitted to the hospital. Traditional approaches are not capable to account for the uncertain nature of risk of hospital re-admissions. More problems arise due to large amount of uncertain information. Patients can be at high, medium or low risk of re-admission, and these strata have ill-defined boundaries. We believe that our model that adapts fuzzy regression method will start a novel approach to handle uncertain data, uncertain relationships between health system variables and the risk of re-admission. Because of nature of ill-defined boundaries of risk bands, this approach does allow the clinicians to target individuals at boundaries. Targeting individuals at boundaries and providing them proper care may provide some ability to move patients from high risk to low risk band. In developing this algorithm, we aimed to help potential users to assess the patients for various risk score thresholds and avoid readmission of high risk patients with proper interventions. A model for predicting patients at high risk of re-admission will enable interventions to be targeted before costs have been incurred and health status have deteriorated. A risk score cut off level would flag patients and result in net savings where intervention costs are much higher per patient. Preventing hospital re-admissions is important for patients, and our algorithm may also impact hospital income.

Table of Contents

Acknowledgements

I would like to thank all people who have helped, supported and inspired me during my doctoral study. Especially, I would like to express my sincere gratitude to my supervisor Prof. Thierry J. Chaussalet for the continuous support of my Ph.D study and research. I am indebted to him for his patience, motivation, enthusiasm, and immense knowledge. His guidance has helped me in all the time of research and writing of this thesis.

I am also very grateful to my Co-supervisor Dr Elia El-Darzi for his crtitical reviews during my second half of my Ph.D study. These reviews were really helpful in improving my study. I would also express my gratitude to Co-supervisor Dr. Panagiotis Chountas for his expert opining on data mining techniques.

My sincere thanks goes to HSCMG (Health Care and Social Modelling Group) for their sincere support in collecting, and understanding of Hospital Episode Statistics (HES) dataset.

My sincere thanks goes to Prof Taj Keshavarz for organizing things and his expert opinion during my Ph.D study.

I would like to thank my husband Rajul Verma, and parents for their consistent support during my entire study. Without their patience, continuous support and encouragement, this work would never have been possible.

Author's Declaration

I, **MANISHA RATHI** declare that this thesis and the work presented in it are my own and has been generated by me as the result of my own original research. I declare that I am the author of the doctoral thesis entitled

"A Novel Framework for Predicting Patients at Risk of Readmission"

List of Figures

List of Tables

Chapter 1

1. Introduction

Hospital management has undergone great changes over the past 20 years. This has led to substantial shifts in demand for hospital care facilities and notable changes in the type of facilities required. The distribution of healthcare use across a population tends to be highly skewed, with small number of people accounting for health care resources (Nuffield Trust, 2011).Admission of a patient is a costly event, and a patient who is frequently admitted could be classified as a high risk patient. Health care managers are facing a set of challenges for e.g ageing and chronic illness are becoming more prevalent, budgets becoming increasingly tight, and a relatively small number of high risk patients accounts for a large fraction of healthcare costs. NHS has recognized that an increasing number of patients are being readmitted to hospitals soon after their discharge. Literature suggests that a small number of patients could be classified as 'high risk' and these patients end up using large amount of hospital resources (Billings et al., 2006; Billings et al. 2012; Billings et al., 2013; Lewis, 2015).Patients at high risk of readmission accounts for high cost in future.

If,these high cost patients could be identified earlier and offered better support and preventive care that might be possible to improve their health outcomes and experience of care. Emergency readmissions are rising in England and many other countries. Unplanned hospital readmissions have been considered as a marker of poor health system performance. One of the fundamental mechanisms underlying hospital re-admissions is their definitions in literature. One of the definitions of re-admission is the number of patients who experience unplanned re-admission within 30 days of the initial admission (Billings et al., 2013). Another definition is the number of patients identified at high risk of readmission within the next 12 months (Billings et al., 2006). Department of Health (DH) in England have provided guidance for restricting payments for readmissions within 30 days of discharge from a previous readmission (Blunt et al., 2014). This policy for non-payment is based on the idea that readmissions are preventable. Unplanned hospital readmissions are common, expensive and often preventable (Gruneir et al., 2011). Many

hospitals lack practical tools to identify patients at risk of unplanned readmission (Bradley, et al., 2012; Bradley et al., 2013).

Multiple hospital admissions represent a particular challenge to health care sector involved in identification of effective methods for hospital resource management (Corrigan and Martin, 1992). If, these high cost patients could be identified earlier and offered better support and preventive care then it might be possible to improve their health outcomes and experience of care (Georghiou et al., 2013).

Risk stratification can also be used to assess the future utilisation of hospital resources by patients, and therefore can aid in the planning of healthcare resources (Thomson et.al., 2013). An approach of risk stratification into high, medium or low is adapted in our framework. This stratification can be uncertain. Risk stratification can assist healthcare professionals in identifying individuals who are likely to be high service users (Adrion et al., 2015). Providing better care to high risk individuals can aid in making large net savings for the health service as a whole (Adrion et al., 2015). There are clear advantages in reducing unnecessary readmissions to the NHS. Researchers have undertaken in-depth research into emergency readmissins and potential financial impact on healthcare decisions. Readmission are widely seen as a problem and in UK, as readmission rates are considered as one of the indicators of quality of care. Readmission to patients are both common and costly, evidence on strategies adopted by hospitals to avoid readmission is limited. Higher readmission rates are associated with lower patient statisfaction and are estimated to cost NHS billions per year in hospital payments (NHS England, 2015). Given these demanding circumstances, health care mamagers are naturally attracted to any intitative that improves that quality of care while simultaneously reducing overall costs (Lewis, 2015; Georghiou et al., 2013). Majority of hospitals reported having objectives to reduce readmission, quality improvement teams focused on readmissions. Specific practices considered to be important for for preventing readmissions were implemented by fewer hospitals.

Mutiple factors increase the chance of readmission for patients discharged from hospital. Several studies have also suggested that patient characteristics such as age, sex, medical history and comorbidity are correlated with early readmission. Researchers also investigates that the quality of care does affect the risk of readmisson with 30 days of

discharge (Ashton et. al., 1996). Unplanned hospital readmissions are more likely to be a possible marker for quality of care. Readmission happens to patients within a total healthcare system, involving care in hospitals, primary care and care at home. Therefore, it is important for health care managers to work with patients and their representatives to audit early unplanned readmissions to improve the quality of patient care

According to previous studies, logistic regression and classification & regression trees have been developed to identify patients at high risk of re-admission. In fact, due to uncertain nature of binary observations, probability distribution cannot be always considered for these types of data. Current probbaility distribution methods may not be appropriate, as it cannot handle range of values for risk of readmission (high, medium or low). Due to ill-defined boundaries of risk of readmission, as patients may move from high to medium and medium to low risk of readmission logistic regression may not be useful.

Research data includes large amount of imprecise observations. More problems will arise when there is an ambiguity in the degree to which an event occurs especially when the relationship between explanatory& response variable are uncertain. (Dom et al., 2008; Shapiro, 2005; Rosma, et al., 2008).

Traditional approaches are not capable to account for the complex action of uncertainty in risk of hospital re-admissions (Coppi, 2008). Three basic sources of uncertainty are considered:

1. Uncertainty in the relationship between response and explanatory variables.
2. Uncertainty about the relationship between the observed data and the universe of possible data.
3. Uncertainty in the observed value of the variables (Coppi, 2008).

There are attempts to construct a fuzzy regression model based on the possibility of success. These possibilities can be defined in linguistic terms as high, medium or low risk of re-admission. However, the borderlines of stratification (high, medium or low) are not crisp and number of readmitted patients' near the borderlines is uncertain. Additionally, relationship between variables is uncertain and it is not modeled in traditional methods. This uncertain nature of re-admission and uncertain relationship cause other difficulties.

As risk of re-admission is uncertain and knowledge about patient's re-admission risk of readmission is imprecise, fuzzy modelling techniques provide a good concept for dealing with such type of uncertain information.(Bisserier et al., 2010) states that fuzzy regression, a type of conventional regression analysis, has been proposed to evaluate the functional relationship between independent and dependent variables in a fuzzy environment (Bisserier et al., 2010).In the present study, the possibilistic approach with a new, revisited methodology is proposed for predicting risk of re-admission of a patient.

The main aim of this thesis is to provide a framework for predicting patients at high risk of re-admission. This framework is depicted in a model, and this model is implemented in an algorithm. Our proposed algorithm adapts fuzzy regression method to predict likelihood of patients at risk of re-admission.

Figure 1 shows the map of the thesis and interdependence among various chapters. In chapter 1 and chapter 2, the background and literature review for our thesis is described. Research aim and research objectives are described in chapter 1.

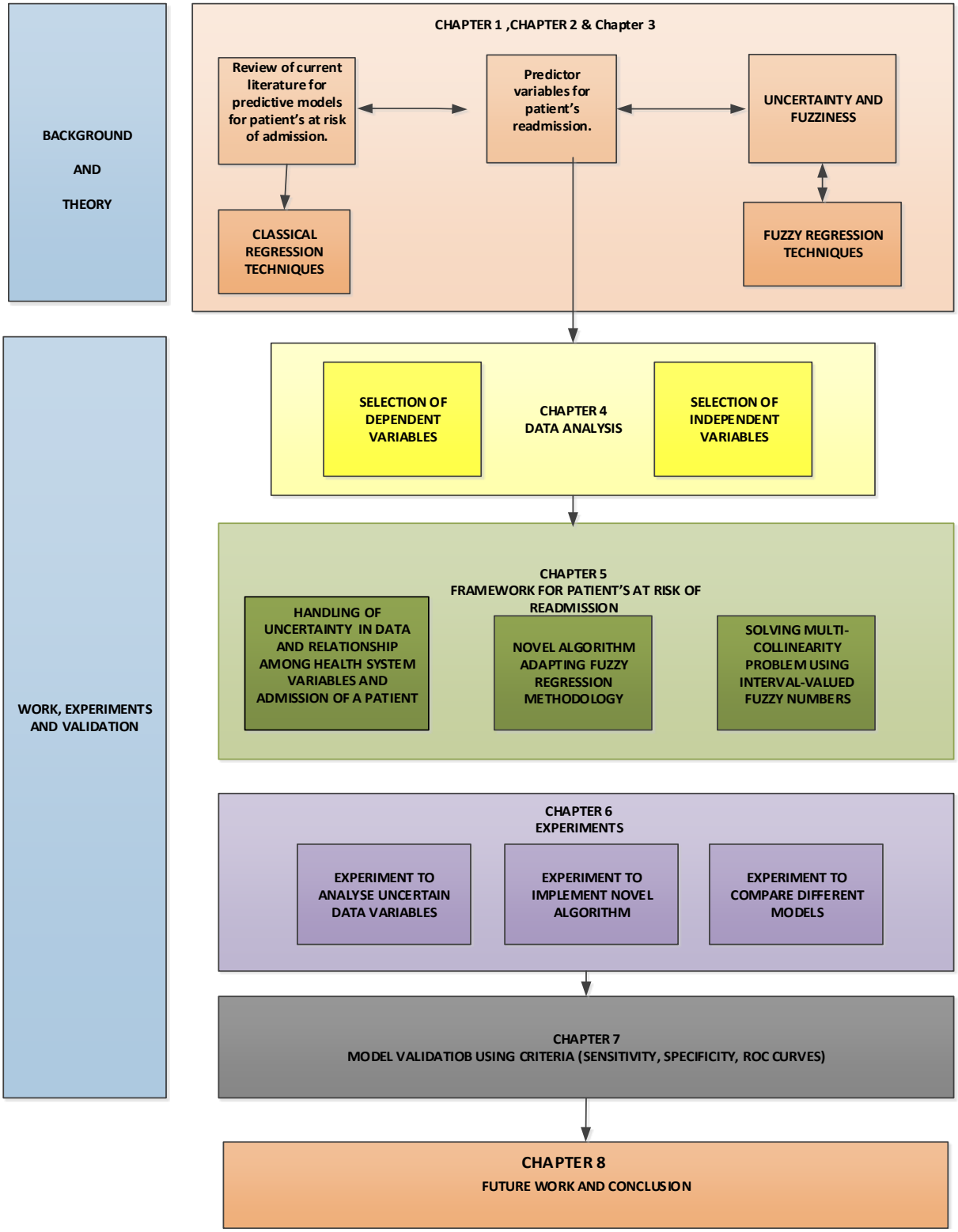

Figure 1 Map and interdependence among various sections of thesis.

The background and literature review chapter describes and reviews current methodologies based on literature. Current literature describes classical regression techniques, fuzzy regression methods and fuzzy techniques in predictive models. Our model is based on the theory of fuzzy regression methods described in current literature. Fuzzy regression methods are used to handle uncertainty in health system variables and uncertain relationship between risk factors and risk of admission. There may be a problem of multi-collinearity in health system variables which could be solved by fuzzy regression method. Theoretical study of fuzzy regression method is described with background, advantages and limitations in chapter 4.

Predictor variables for patient's re-admission are identified after studying various articles. Risk factors for a patient's re-admission are also studied in the literature review. Risk scores are used to stratify risk of readmission into high, medium or low. Therefore, risk scores of different models are compared and discussed in literature review. An approach of risk stratification into high, medium and low risk is adapted to identify high-risk individuals and enable proper interventions for high-risk patients. Because our model's performance is compared and evaluated with other methods, a comparison of various models using performance evaluation methods is done in literature review.

Once predictor variables are identified, a conceptual framework for identifying patients at risk of re-admission is developed as in chapter 6 and implemented using our novel algorithm. This novel algorithm adapts fuzzy regression method for predicting patients at risk of re-admission and identifying risk factors responsible for likelihood of re-admission. This novel algorithm handles uncertainty in risk of re-admission. Also, the algorithm could be further extended to handle uncertainty and multi-collinearity problem within health system variables.

Our algorithm as described in chapter 6 is experimented in chapter 7. Chapter 7 consists of different sets of experiments where we have handled uncertain data variables compared different models, and implemented fuzzy regression method. Fuzzy membership function for risk of re-admission is also described in the experiments. For our research work, we have focused on triangular and trapezoidal membership function. For the experiments, health system data variables are analysed. Data analysis includes steps

such as selection of independent & dependent variables, fuzzification of response variable, outlier detection and handling multi-collinearity problem. Model validation is done in chapter 7, where different models are compared and evaluated using model calibration and discrimination techniques. Models involved in comparison with fuzzy regression method are logistic regression, decision tree, and neural network methods.

Conclusion of overall thesis is given in chapter 9. Research limitation and novel elements of the research are described in this chapter. The limitations of the research are identified and future work is proposed to address those limitations.

1.1 Overview of Risk Prediction Models

According to literature, different authors have explored the predictive power of various types of model for risk prediction. These papers have sought to use models for different outcomes (for example hospitalisation, cost etc.) which makes comparison difficult. For instance, Adjusted Clinical Groups (ACGs) and Diagnostic Clinical Groups (DCGs) based on age, gender, and diagnoses were designed to predict future costs for the individuals in need of hospital resources. ACGs and DCGs uses ICD9-CM diagnosis to classify patients with special attention paid to individuals with expensive chronic conditions (Johns Hopkins ACG, 2014). Our focus is on a range of different models, which have been used in NHS to identify people at high risk of re-admission to hospital (Billings et al., 2013). We have studied models used for risk prediction over past years in the UK such as Patient at Risk of Re-admission (PARR), PARR++, Combined Predictive Model (CPM) and ACGs. These models are used for predicting events such as unplanned hospital re-admissions which are undesirable, costly and potentially preventable (Billings et al., 2006, Billings et al., 2013). Predictive models work by combining information at patient level to identify potential co-variates that are associated with a future event- such as likelihood of re-admission to hospital.

ACG-based models were specifically calibrated to identify patients with risk of future hospitalization. ACG based models focus on unanticipated hospitalization, and are used to estimate future resource utilization for sub-groups within a population (Johns Hopkins ACG, 2014). The ACG system is a suite of tools which draw on demographic, diagnostic, pharmacy and utilisation data from primary and secondary care. Currently, there are

different kind of models use for risk prediction for e.g PARR, CPM and ACGs. ACG models were developed based on US healthcare data. For ACGs system developed in John Hopkins University licensing arrangement is required. ACGs system predicts likelihood of readmission and likely cost of a patient in the coming year (Johns Hopkins ACG, 2014).

Of the various models in use in England the PARR tool has gained popularity. This is probably because of the data to run PARR is easily available and software to estimate risk scores was distributed evenly. PARR1 and PARR2 tools that identify high risk patients use inpatient data to produce a 'risk score' showing a patient's likelihood of re-hospitalisation within the next 12 months (Billings et al., 2006, Corrigan and Martin, 1992) and (King's Fund , 2006; Foot et al., 2014) developed a number of predictive tools for the prediction of patients who are at high risk of re-admissions. (King's Fund , 2006) developed these stratification models:

- The Patients at Risk of Re-hospitalisation (PARR1) tool: a software tool that uses inpatient data on prior hospitalisations for certain 'reference conditions' to identify patients at risk of re-hospitalisation within a year.
- The Patients at Risk of Re-hospitalisation (PARR2) tool: a software tool that uses data on any prior hospitalization to predict risk of re-hospitalization. It extracts information from Hospital Episode Statistics (HES) data using criteria that are known to be risk factors in future admissions to hospital.
- The Combined Predictive Model: a model that uses inpatient, outpatient, Accident & Emergency (A&E) and General Practitioner (GP) data to stratify populations according to their risk of admission. The combined model takes primary and secondary care data for entire patient population and stratifies those patients based upon their risk of emergency admission in the next 12 months (King's Fund , 2006).

Such tools can be used to identify patients for an appropriate intervention in order to improve health outcomes, and to allocate resources efficiently. If the system is able to identify those patients who are at the highest risk of re-admission, more intensive resources can be focused on them leading to efficient allocation of resources and facilitating better planning of services (King's Fund , 2006; NHS England:, 2015).The stratification of patients can be illustrated using Kaiser Permanente's triangle below in

figure 2. At the top of the triangle are the individuals who are most at risk of emergency admission. Predictive models attempt to target individuals at the top of triangle to prevent them from being readmitted. However, decision makers are not sure whether this is the most appropriate area of triangle on which to concentrate resources. It may be useful to identify individuals in the lower two strata who are likely to move into the high risk level. The borderlines of these risk stratifications are not crisp, and individuals near the boundaries can move from low level to high level. Evidence is weak as to which strata an intervention is applied to get best results. Existing predictive models seek to establish relationships between set of variables in order to predict future outcomes. Most predictive models that focus on regression techniques are able to identify patients at risk of readmission (Billings et al., 2006, Bottle et al., 2006, Billings et al., 2013). As an alternative to regression, researchers have applied various machine learning methods especially Artificial Neural Network (ANN) and support vector machines (SVMs). Initial results from these have been promising (Bottle et al., 2006; Bottle, et al., 2014). However, the users of these models are unable to know how exactly these models predict risk and thus the relationships between inputs and outputs. In contrast, decision tree methods show the relation between predictor variables. As the status of the individuals near the borderlines of risk stratification triangle is uncertain, we have adapted fuzzy regression method in our research, which is shown in detail in later chapters.

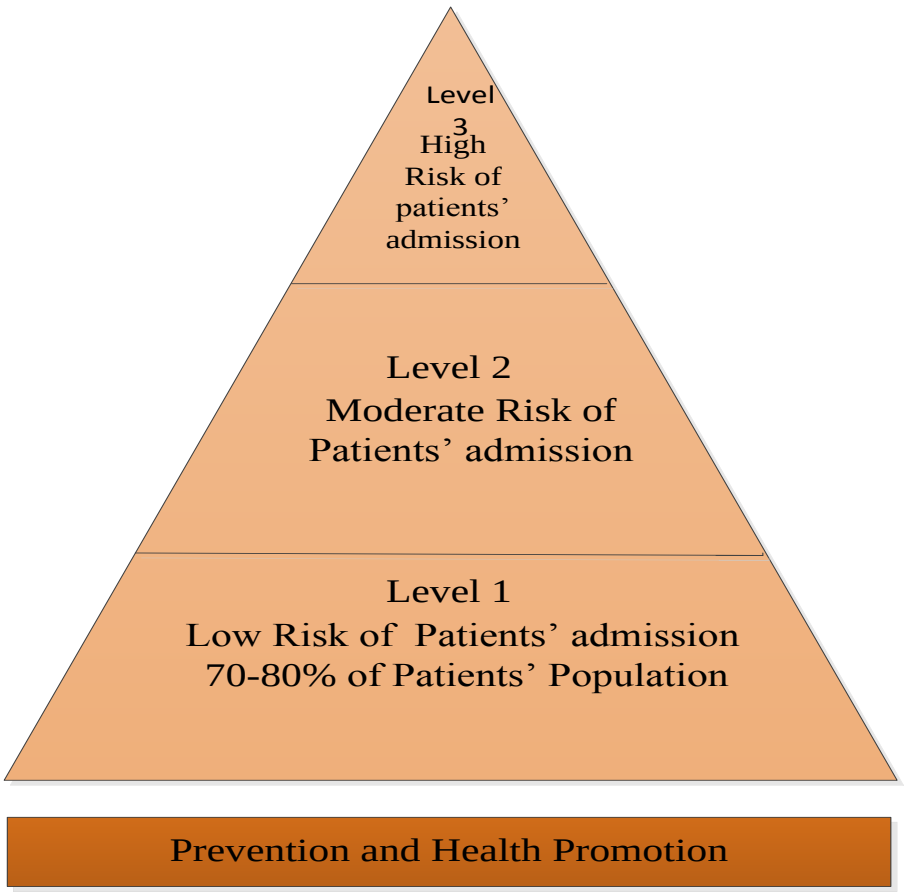

Figure 2 Risk stratification triangle as developed by Kaiser Permanente (King's Fund, 2006).

1.2 Research Aim

The aim of this research is **"To develop a novel framework for handling uncertainty and fuzziness in predictive models for health care services, to enable efficient health care resource utilization"**.

Research Objectives

The specific objectives of the research are:

- To develop a framework that identifies individuals at risk of readmission within 12months of discharge. This framework will identify significant predictors for risk of re-admission, and identify patients at high risk of readmission.
- To develop a model which is implemented in a novel algorithm. This novel algorithm will capture the uncertain nature of risk of readmission and stratifies patients at high, medium or low risk. It will adapt fuzzy techniques (Fuzzy linear

regression), so that it can handle uncertainty in risk of readmission. Additionally, the algorithm will also handle problem of multi-collinearity that occurs among health system variable.

- To produce a risk score from this model for readmitted patients, and percentage of patients with a re-admission within 12 month of discharge period in different risk bands. To measure the area under the receiver operating characteristic curve, together with Positive Predictive Value (PPV) and sensitivity for a range of risk thresholds.

1.3 Research Motivation

Risk of re-admission of a patient can be viewed as a fuzzy event because it can take values other than 0 and 1, and also unplanned re-admissions do not have clear cut boundaries. The principal of stratifying patients according to risk is relevant and useful in order to improve health outcomes and to facilitate better planning of resources. Tradtional model such as PARR used statistical techniques such as logistic regression to predict future outcomes.

In statistics, linear regression analysis is a powerful method for studying the linear relationship between one response variable Y (dependent variable or output variable) and a set of explanatory variables $X_1,,,,,,,,X_P$ (independent variables or input variables) (Dirusso et al., 2002).Classical statistical regression has many applications, problems may occur in some situations. Linear Regression method is extremely sensitive to outliers. Other situations are:

- Imprecise Information
- Uncertain data
- Vagueness in the relationship between input and output variables.

In recent years, there is a growing literature that formalizes the linear regression model in fuzzy domain, in which model parameters and/or data are fuzzy, or imprecise or vague (Dirusso et al., 2002). Abundance of vague observations in healthcare studies, motivate us to think about a proper model in a fuzzy environment. These are the situations fuzzy regression was meant to address.

1.3.1 Uncertainty

Any decision making statement is either true or false. Uncertainty in decision-making arises due to the non-uniform data and lack of knowledge in data characteristics. Due to lack of knowledge, we can only estimate to which degree they are true or false. In our research context, uncertainty can be defined with respect to risk of re-admission of patients.

1.3.2 Vagueness

Vagueness theory can account for all those approaches in which statements (such as “risk of re-admission is high”) are true to some degree. Stratification of risk of re-admission can be done in linguistic terms such as high, medium or low risk of re-admission, but these have ill-defined boundaries. These ill-defined boundaries of linguistic categories can be referred to as vagueness and can be captured with membership functions in fuzzy sets.

1.3.3 Imprecision

(Bosc, 1995) Null set denotes the lack of information about a value. At times, it is known that a missing value belongs to a more limited set of values (possibly, a range of values), which are known as disjunctive values (Motro, 1995). Null and disjunctive values both express imprecision.

1.4 Research Contribution

The main contribution of this work is both conceptual and practical. The contribution refers to the development of a conceptual framework that adapts novel approach for the prediction of response variable based on fuzzy regression method. Patients can be stratified into risk bands, which range from high to low depending on their risk scores. Percentage of patients re-admitted in a high risk band can be evaluated with the help of risk score threshold value. Health care and social interventions could be better targeted at individuals who are at high risk of re-admission and most in need of hospital resources.

The research is novel in the sense that it is the first study that handles linguistic variable risk of re-admission, which can be defined by a fuzzy set with range of values as high, medium or low risk. The methodology proposed here is original in the sense of designing and developing a framework that models uncertainty in risk of re-admission of a patient.

Traditional methods of prediction such as logistic regression are not able to account for uncertain nature of risk of hospital re-admissions.

Our model is based on a framework that identifies significant predictor variables as risk factors to predict patients at risk of admission. The knowledge of the impact of risk factors will provide clinicians better decision-making and will help in reducing the number of patients re-admitted to the hospital.

The other important contribution is the design and development of a novel algorithm that adapts fuzzy regression method. This algorithm deals with uncertain data and uncertain relationships between risk factors and risk of re-admission. It estimates the unknown dependency between the independent health system variables and the response variable. We believe that this will start a novel approach to handle uncertain data, and uncertain relationships between health system variables and the risk of re-admission using the possibility approach with revisited methodology.

1.5 Research Impact

Various risk stratification models aimed at identifying individuals at risk of hospital readmission have been developed, using healthcare data. Risk stratification tools are designed to identify those individuals who are at high risk of experiencing an adverse future outcome, such as readmission with 12 months or 30 days of discharge. It could be beneficial to identify these high risk individuals and provide interventions to reduce hospital readmissions. We will identify patients at high, medium or low risk with the help of risk scores. A higher risk score will imply higher probability of future re-admission than lower risk score. This could lead to net savings in costs incurred for hospital services. Such an approach relies on the ability to identify appropriate patients. Targeting individuals at boundaries and providing preventive care to such patients may help in proper utilization of hospital resources. Risk stratification tool is ever completely accurate, and there are very likely chances that high risk indivuals may move to medium or low risk bands. Borderlines of risk stratification are not crisp, and boundaries for high, medium and low risk of readmission is not clearly defined. This may lead to poor clinical decision making, and hospital resources may not be properly utilized.

High costs may be incurred for treating patients at boundaries, which are ill-defined, and strategies designed to improve the impact of treatment could worsen consumption of health care facilities. Some of the individuals may face a problem that they are offered an intervention to prevent an event which they were actually not going to experience. They might go through over treatment, which may increase their anxiety levels and result in unnecessary side effects. On the other hand, patients who are actually in need of resources might not receive any these services. Effective risk stratification tool is the one where benefits to the population outweigh the costs for hospital services.

Patients at risk of re-admission could be identified with consideration of significant variables. A model for predicting patients at high risk of re-admission will enable interventions to be targeted before costs get incurred and health status gets deteriorated. In developing this algorithm, we aim to help potential users to assess the patients at various risk score thresholds and design proper interventions for individuals who are at high risk. A risk score cut off level would flag patients at high risk and where intervention costs are much higher per patients. Because of nature and ill-defined boundaries (high, medium or low) of risk of readmission this approach does allow the user some ability to compensate for number of patients at high risk of re-admission. Preventing hospital re-admissions is important for patients, and our algorithm may also impact hospital income.

Hospital re-admissions are also used as indicator of quality of care. The validity of hospital re-admissions as an indicator of quality of care depends on the extent that hospital re-admissions are avoidable.

1.6 Research Challenges

- For our research, Hospital Episode Statistics (HES) data is used. There are about 300 variables in HES data base. Understanding of each and every variable and its impact on hospital readmission is a challenge. To deal with HES variables, we have done study of all available information on HES data. Statistical analysis (chi-square) is done to understand significance of variables in patient's readmission.
- Data collected from Hospital Episode Statistics datasets is challenging to deal due to size and complexity of the dataset. Because of complex coding of data items,

missing data, duplicates and other data issues lot of analysis is required to produce meaningful information that are free from errors.

- Due to difficulty in handling of dataset and license problem in MATLAB, fuzzy regression method was difficult to implement in MATLAB. R was used because it was more user friendly with easily available open source package for implementation of fuzzy regression method.
- Because of non-uniform data and lack of information about data, there is difficulty in making sharp and clear distinctions in the real world. Risk of re-admission can be defined in linguistic terms which result in ill-defined boundaries. This can be solved by representing risk of re-admission by fuzzy membership function.
- Missing and inaccurate data (coding in diagnostic fields) was a problem, as model is dependent on these fields. These data limitations tend to give error in prediction of the model.

Chapter 2

2. Literature Review

2.1 Introduction

(Lewis, 2015)states that the number of unplanned re-admissions in the UK National Health Service (NHS) and in several other developed countries' hospitals has been rising for many years. The NHS in England is the publicly funded healthcare system for England.It believes in the long-held ideal that good healthcare should be available to all, regardless of wealth (NHS England, 2015). It deals with over 1 million patients in every 36 hours. An essential strategy of NHS is to provide is to improve care and services for high cost patients.

(Hasan et al., 2010) states that studies about hospital re-admissions have focussed on specific conditions or populations and generated complex prediction models.

Our focus is on range of risk prediction models, which have been used in NHS to identify people at high risk of re-admission to hospital. The predictive model seeks to establish relationships between sets of variables in order to identify patients at risk of re-admission (Billings et al., 2013). (Curry et al., 2005; NHS England,2015; Lewis, 2015; Purdy, 2010) forecasts risk of future event based on the identified relationships between nunber of factors and increased rates of readmission.

Risk of re-admission is an uncertain event which has important economic implications for efficient hospital resource utilization. This health care problem deserves modelling of all relevant uncertain information involved in the real decision-making process (Hojati et al., 2005, Pourahmad et al., 2011).The literature review consists of various sections. Sub section reviews the literature on predicting unplanned hospital resource utilization, predictors for risk of re-hospitalisation, data mining and machine learning techniques and fuzzy methods. Data mining and machine learning technques and fuzzy methods are included in chapter 3.

2.2 Methods

2.2.1 Literature Search

We developed a search strategy to identify studies that measured the proportion of re-admission deemed avoidable. We searched databases such as pubmed, medline, web of science etc for papers published from 1965 to 2015. Full text versions of citations were retrieved for complete review. The references of all included studies were reviewed to identify other eligible analyses. Data abstracted from each study included basic study information (publication, year, and journal). This study is not without limitations. Although, we did a comprehensive search of the peer reviewed literature, we did not include unpublished results or studies from the grey literature such as reports and doctoral dissertations. Some of the reports included in our research are from Nuttfield Trust, and Health Services and delivery research which are relevant for our thesis.

2.2.2 Study Selection

Search strategy for the systematic review was conducted in the electronic database from 1965 till 2015. Keywords used were “patient re-admission”, “hospital re-admission”, “risk of re-admission”, “risk stratification”, “unplanned admission”,” uncertainty”, “chronic illness”, “rehospitalisation”, “risk factors”, “Data mining”, “Machine learning”, “Fuzzy methods”, “fuzzy regression”, “Machine learning Algorithm”,” predictive modelling. In all searches a filter was used for systematic review. Systematic reviews were hand searched. Articles were independently screened for titles and abstracts for inclusion. Exclusion criteria were unpublished abstracts and dissertations. The abstract of each article was reviewed for inclusion in the sample. In situations when inclusion could not be determined by abstract review then full text articles were reviewed. Full text copies of most of the potentially relevant papers were retrieved and checked formally for eligibility. Thus, the final sample included articles reporting studies of hospital re-admissions.

2.3 Unplanned re-admissions and hospital resource utilization.

Staff and managers in hospitals and other health care settings are under pressure and are concerned for effective use and management of scarce resources (Chan et al., 2011). Unplanned hospital admission is increasingly recognized as a significant contributor to rising health care costs (Pourahmad et al., 2011). Some of the studies for predicting unplanned re-admissions are mentioned in Table 1 of Appendix 1. Health services are a vital part of the NHS for millions of people, and they comprise approximately £10 billion of NHS budget (Foot et al., 2014). For several years, the NHS has recognized that an increasing number of patients are being readmitted to the hospital. Patient's re-admissions to hospitals are associated with increased costs to the hospitals. Re-admissions to hospitals are being used as indicator of quality of care (Briefing NHS Confederation, 2011). Health care services collect and analyse detailed readmission data to understand disease, clinical practices, patient characteristics and factors driving readmission trends (Sg2, 2011) One of the fundamental mechanisms underlying hospital re-admissions is their definitions in literature (Billings et al., 2013; Bottle et al., 2014; Kansagara et al., 2013). Identifying patients at risk of readmission within 12 months is selected as the time period of 12 months may give clinicians and healthcare managers to contact and high risk patients. It also allows time to initiate behavioural and treatment changes. Selection of shorter time frame may improve the accuracy of the prediction but decrease usefulness of prediction (Au et al., 2012).

2.3.1 Hospital re-admissions as an indicator of quality of care.

Information about care is important as it helps us to assess how health and social services can be co-ordinated in ways to provide higher quality and more efficient care. Hospital readmissions are indicator of quality of care (Briefing NHS Confederation, 2011; van Walraven et al., 2011). In many healthcare systems, the unplanned readmission to a hospital has become indicator of quality of care which measures how many patients are readmitted to hospitals after they have been discharged (Kossovsky et al., 2000). Implementing the methods to provide interventions for patients at risk of readmission may help end variations in quality of care and finances that cost NHS billions (Appleby et al.,

2012; Department of Health, 2012). Yet, there is a shortage of information about the care that people receive at the time of re-admission. High quality patient care and sound financial management go hand in hand (Department of Health, 2012). The validity of hospital readmissions depends on the extents that they are avoidable (Purdy, 2010). As the number of avoidable hospital readmissions increases, the cost required to avoid one readmission will decrease. Some these readmissions are avoidable, while at other times they are unavoidable due to the development of new conditions or severe chronic conditions (Purdy, 2010). In most case, these unplanned hospital readmissions indicate bad health outcomes for patients. Unplanned hospital readmissions are a problem for health care systems as they are costly and lead to bad quality of care (Briefing NHS Confederation, 2011). (Sg2, 2011) estimates that total penalties associated with 30 days of emergency readmissions would potentially cost NHS trusts £584 millions in lost income. Therefore, many healthcare organizations use risk prediction models to target interventions aimed at preventing hospital readmissions.

The trend of unplanned admissions, which is possibly related to poor patient care, places financial pressures on hospitals and on nation health care budgets (Sharon et al., 2004; Friedmann et al., 2001; Hensher et al., 1999). The majority of studies in the papers reviewed look at risk of unplanned hospital re-admissions (Parker et al., 2003). The main aim of study is to identify people at risk of re-admission whose health outcomes may improve by direct use of intensive resources. The issue is that a small number of patients could be classified as high risk patients who are using a large amount of resources (King's Fund, 2006; Georghiou, et al., 2013; Lewis, 2015). The identification of readmitted patients may also provide information of the severity of condition and the quality of care provided for them. The focus is on identifying 'high risk' patients for whom an appropriate intervention would improve care and prevent future re-admissions. Assessment of the selected predictive variables may help support the development of plans that potentially mitigate the risk of re-admission (Fialho et al., 2012).Therefore, studies for identifying predictors for hospital re-admissions are considered.

2.4 Predictors of hospital re-admissions

Collection of the predicted variables is not so complex in hospitals and their assessment may help in support of the development of healthcare management plans that potentially mitigate the risk of re-admission. Studies for identifying potential predictors is mentioned in next section, and significant predictors are listed in Table 1. Previous studies have examined different variables that are assessed at discharge and that are considered to be predictive of re-admission (Fialho et al., 2012) as shown in Table 1. Some of the variables included in PARR case finding algorithm are age 65-74, age 75+, sex, ethnicity, previous admission for a reference condition, number of emergency admission in previous 90,180 and 365 days, number of non-emergency admissions in previous 365 days, average number of episodes per spell for emergency admission, and diagnostic cost groups/hierarchical condition category (Curry et al., 2005; Au, et al., 2012). Table 1 lists predictor variables and their corresponding authors and published papers. Variables selected for the highest predictive power include patient, demographic characteristics and health system variables. These variables are selected on the basis of our hypothesis that a good predictive value can be achieved, and the variables chosen should be independent of correlation. We have done a comparison of predictor variables. We have selected final model variables as severity illness, prior hospitalization, comorbidity, overall demographic and patient characteristics and health system variables. (Billings et al., 2006, Billings et al., 2012, Billings et al., 2013) created a set of variables on previous hospital use and diagnostic history for hospital episode statistics data for triggering admission.

Several risk factors are also assessed for likelihood of re-hospitalization and are considered to be significant for predicting high risk patients Current literature has examined risk factors for unplanned re-admission that relate to health care factors, patient characteristics such as age, gender, home living situation and stage of illness, factors related to the patient or a combination of all these (Bisserier et al., 2010; Chan et al., 2011). The various risk factors for patients' re-admission are shown in Table 2 of Appendix 1. A number of factors are associated with increased rates of readmission. Age is one of the potential predictors of readmission (Purdy, 2010; Philbin et al., 1999; Gruneir et al., 2011). Data on the impact of ethnicity on readmission is limited. Being from minority ethnic group is associated with a higher risk of readmission (Bottle et al., 2006; Purdy, 2010). For each

of the predictive models, patient comorbidities were identified using ICD-9 and ICD-10 codes from the hospitalizations in the past 12 months. Charlson index was computed using ICD-10 codes (Au et al., 2012; Aylin et al., 2010). Higher leveles of morbidity in population are associated with higher levels of readmission, and readmission rates are correlated with chronic illness(Bottle, et al., 2006; Purdy, 2010).

Predictors	(Philbin and Disalvo, 1999)	(Howell et al., 2009)	(Kansagara et al., 2011)	(Billings et al., 2006b)	(Lagoe Ronald et al., 2012)	(Gruneir et al., 2011)	(Dorman et al., 2012)	(Vilaro, et al., 2010)	(Graham et al., 2013)	(Pavon et al., 2014)	(Kasper et al., 2002)
Age(yr)+		√	√	√	√	√	√	√	√	√	
Gender		√	√		√	√		√	√	√	
Height/ Weight (BMI)							√				
Indigenous status		√									
Marital status		√	√					√			
Comorbid illness	√	√	√	√		√	√			√	
Charlson comorbidity index	√		√	√							
Mental health comorbidity			√								
Ethnic origin	√			√		√				√	
Insurance	√										
Hospital Location type	√	√			√				√		
Index of Multiple deprivation band for the place of residence				√							
Discharge to a skilled care	√				√						
Previous hospital discharge	√			√							

Table 1 Predictor variables for risk of re-admission.

2.5 Summary

In this chapter, we found the literature on unplanned admissions. We considered a substantial amount of published literature and found number of studies that met the eligibility criteria. Currently, an available published study shows literature on unplanned admission, predictors for hospital re-admission and hospital readmissions as indicator of quality of care. We conducted a systematic literature review of studies that measured the proportion of re-admission that are avoidable.Previous studies have examined different variables that are assessed at discharge, and risk factors that are considered to be predictive of re-admission. In the next chapter, we have examined studies for data mining and mahine learning methods.

Chapter 3

3. Data Mining and Machine Learning Methods

3.1 Introduction

Literature suggests a number of different regression models in order to target patients at high risk of re-admission (Billings et al., 2013;Demir, 2014).The outcome of linear regression is the actual value whereas logistic regression produces a predicted probability between 0 or 1 for an event, such as admission(Zhao et al., 2003, Meenan et al., 2003).Use of either method is valid, and logistic regression can only be used as long as the variables are appropriately transformed in order to build such a model. Recently, data mining and machine learning techniques are used for developing models to solve healthcare problems. Data mining algorithms are applied in variety of healthcare problems. (Liu et al., 2006) applied data mining algorithms to predict inpatient length of stay in geriatric hospital treatment. Literature on application of data mining andmachine learning algorithms in various healthcare domain areas is studied.(Zernikow et al., 1999) studied the accuracy of two LOS prediction models, namely a multiple linear regression model (MR) and an artificial neural network (ANN) for outlier detection in the hospital admission. More literature on data mining and machine learning algorithms is included in chapter 3.

There are number of studies that compare the performance of regression trees and logistic regression for predicting outcomes. (Austin, 2007; Bottle, et al., 2014) compared the performance of logistic regression method with ANNs, Support vector machines (SVMs) and decision trees. The machine learning methods tend to be slower than Logistic regression (LR), given their complexities, often need an expert as an operator to make decision on implementation issues (Bottle et al., 2006; Demir, 2014). In this section, we have defined machine learning methods such as artificial neural network (ANN), logistic regression (LR), neuro-fuzzy, and fuzzy regression methods in areas of healthcare research.

3.2 Logistic Regression methods

There are attempts to create risk prediction models using descriptive statistics including univariate and multivariate analysis, and validating these models on a sample of dataset. Basic descriptive statistics were conducted on prediction techniques.

In one of the studies by (Rasmusson et al., 2013)a multivariate analysis was performed to predict heart failure re-admissions. Univariate logistic regression analyses were performed to investigate factors that were significantly associated with increased risk of re-admission(Moran et al., 2013).. Multi-variable logistic regression analyses were performed using a stepwise selection of variables to evaluate for statistical differences between those patients with re-admission and those without re-admissions.(Su et al., 2013)applied cox-regression on the data to solve the prognosis view of the re-admission risk prediction problem.(Kansagara et al., 2011) reported the c-statistic, with 95% confidence interval to describe the model discrimination. The c-statistic, which is equivalent to the area under the receiving operating characteristic curve, is the proportion of times the model correctly discriminates a pair of high and low risk individuals.

Logistic regression is a widely used statistical method in healthcare research (Concato et al., 2001)Logistic regression is a kind of generalized linear regression model that is widely used for prediction of the probability of occurrence of an event(Lin et al., 2010).

We have chosen a subset of logistic regression for illustrative purpose. We are interested in identifying those variables that contributed most to the predictions of readmissions within 30-day discharge. Logistic regression models were constructed to identify such variables, and estimating probability of readmission within 30 days by creating risk score ranging from 0.01-1.00. Significance of variables responsible for risk of readmission was evaluated using statistical tests for e.g pearson correlation coefficient. Logistic regression was considered to calculate the probability of an event given risk factors. We also carried out univariate analysis to identify variables significantly associated with readmission. All variables with value of $p<0.05$ were included in our multiple regression model. As we have large dataset of potential independent variables for our analysis, and inclusion of all variables may decrease the precision of estimated coefficients and predicted values. Secondly, we also want to include as few variables as possible. Therefore, we followed multivariate logistic regression with backward elimination method to select most potential variables to be included for analysis.

Multi variable analysis using a stepwise logistic regression model was used to identify independent risk factors for a 30-day re-admission (Billings et al., 2006). Demographic, clinical and social variables were obtained at baseline and included in a multivariable

logistic regression analysis to identify predictors of early re-hospitalisation (Muzzarelli et al., 2010). The logistic regression analysis identified independent predictors for re-admission to the Intensive care unit (ICU). Patient factors association with hospital re-admission was fitted with multivariable logistic regression models for each of the patient factors using data from validation dataset (Hasan et al., 2010). Variables are often selected for inclusion in logistic regression models using some form of the backward or forward stepwise regression technique (Billings et al., 2006)

3.3 Artificial Neural networks

In the last decade, the use of data mining techniques has become widely accepted in medical patients (Tu, 1996). Prediction techniques are useful in many areas of healthcare research. Amongst the methods used for outcome prediction, artificial neural networks (ANNs) are powerful tools to use in the simulation of various nonlinear systems and they have been applied to both risk evaluation and prognosis of medical science (Tu, 1996). The ANN models used had three layers, one input layer, one output layer, and one hidden layer. Each layer consists of a set of nodes that stimulate humans' neurons. (Tu, 1996)ANN developed a typical neural network consists of three nodes that are arranged in three layers (input, hidden, output). In a neural network, predicting an outcome is based on the values of some predictor variables. Neural networks can be developed with multiple hidden layers but there is no advantage of doing so. As stated by (Tu, 1996) each node in the input layer is usually connected to each node in the hidden layer, and each node in the hidden layer is connected to each node in the output layer.

ANN models are used to evaluate the relationship between subjective patient QOL (Quality of Life) assessments and QOL assessments made by pharmacists and nurses (Takehira et al., 2011). QOL parameters were modelled with ANN using the scores given by patients regarding health related QOL as input variables.(Tan-Nai Wang et al., 2013) (Wang et al., 2013) trained an ANN model to predict cancer patient's five-year sustainability. The artificial neural network used in this study was a multilayer perceptron (MLP) ANN and the number of hidden neuron was set as ranging from 5to 15. The ANN was developed using the structure of multi-layer perceptron (MLP) with back-propagation (a supervised learning algorithm) (Lundin et al., 1999). A better understanding of the

"hidden" layers in neural networks enables to create models that incorporates the best features of neural networks (Ottenbacher et al., 2004).

3.4 Decision trees

A decision tree is a set of if-then clauses (nodes) with a tree like structure, with each leaf being a decision on the expected outcome (Bottle et al., 2006; Bottle et al., 2014).Classification and regression trees (CART) methods are important at identifying important interactions in the data and in identifying clinical sub groups of subjects at very high or very low risk (Demir, 2014; Auble et al., 2005). (Lemon et al., 2003) has adapted classification trees to identify a subgroup of patients with an observed rate of discharge or serious medical complications before discharge. (Austin, 2007) describes a model with decision tree, which decides the most significant independent variable in each stage of predicting depending variable. An ensemble of single decision trees is known as random forests. Decision tree finds the data features that are most important. In single decision trees we find data features that best splits the data into classes, and is repeated recursively until data has been split into homogeneous groups. While in random forests during learning tree nodes are split using a random subset of data features. In random forests, all single decision trees contribute to produce a final answer (Ali et al., 2012).

Data driven methods, such as classification and regression trees (CART) have been used to identify risk of adverse outcomes (Demir, 2014; Tsein et al., 1998). Classification tree-based models is powerful and is known to be a data-intensive approach that works well with large datasets (Freidmann et al., 2001). One of the important advantages of this approach is that tree-based methods are adept at identifying important interactions between predictor variables.

3.5 Fuzzy neural networks in predictive models

Uncertainty characteristics of data is a challenging problem for modelling (Dom et al., 2008; Zadeh, 2005). In past decades many soft-computing based techniques were proposed. Different levels of hybridization on soft computing techniques were also proposed. Among them fuzzy neural network (FNN) based systems are promising due to their capability of modelling data uncertainties (Chen & Wang, 2012). Fuzzy neural network has gained popularity in medical applications (Steimann, 2001). FNN based systems use

fuzzy rules, whose input, antecedent and consequent fuzzy sets use membership function (MF).

3.6 Fuzzy regression method in predictive model

Fuzzy linear regression may be used as an alternative to statistical linear regression model due to vague relationship among variables and poor model specifications (Kim et al., 1996) Classical regression techniques make rigid assumptions about the statistical properties of the model e.g. the normality of error terms and predictions. These assumptions are difficult to justify unless a sufficiently large dataset is available (Kim et al., 1996) In classical statistical regression model, which uses a linear function to express the relationship between a dependent variable y and the independent variables $x_1, \dots \dots \dots x_n$ the parameters are crisp numbers and the error terms are present due to measurement errors (Kim et al., 1996; Tanaka, 1989).

On the other hand, (Shapiro, 2005) defines fuzzy regression as a method to estimate the deviations between observed and estimated values. Fuzzy regression gives rise to a possibility distribution that account for the imprecise nature or vagueness of our understanding of a phenomenon (Shakouri G & Nadmi, 2009). Fuzzy linear regression is used to express the uncertain relationships between system target values and their characteristics, and the interrelationship among characteristics (Tanaka , 1987).

We consider two cases: First, when only the dependent variable is fuzzy. Secondly, when both dependent and independent variables are fuzzy. The aim of the fuzzy regression approach is to determine the functional relationships that lead to the development of a programming model (Sener et al., 2011;Tanaka, 1989; Ramli, et al., 2011). The fuzzy regression approach determines the spreads and the centre values of the regression parameters to estimate uncertain relationship. H € [0, 1] is represented as measure of goodness of fit and is selected by the decision maker (Hojati, et al., 2005), where H represents the minimum degree of certainty acceptable. The purpose of goodness of fit of a regression model is to know how well a model fits a given set of data, or how well it will predict future set of observations.

3.7 Comparative analysis of different algorithms Literature on comparative analysis of different algorithms.

In section 3.7, a literature review of various data mining and machine learning techniques in healthcare applications is done. A brief of summary of different algorithms with benefits and disadvantages is represented in Table 2.

Author(s)	Algorithm	Type of Dataset	Benefits	Limitations
(Abbod et al., 2007) (Razi & Kuriakose, 2005) (Ramesh et al., 2004) (TU, 1996) (Ottenbacher et al., 2004)	Artificial Neural Network(ANN)	Cancer dataset Paediatric Trauma Patient data set	-Can model high dimensions. -Can model non-linearity. -Functional relationship between Independent and dependent variables are unknown. -Performs well for few attributes.	-Needs a lot of data. -Non-Transparent. Training depends on cost function. -Great computation burden.
(Ottenbacher et al., 2004) (Razi & Kuriakose, 2005) (Dom et al., 2008)(TU, 1996) (DI-Russo, 2002) (Su et al., 2006) (Pourahmad et al., 2011)	Logistic regression	Smoker data set Cancer data set HES data set Paediatric Trauma Patient data set SLE(Systematic LupusErythenalosis) dataset. Stroke Patients data set	-Easy to construct -Ability to perform 'optimal' input variable selection. -Ability to explain relationship between response and input variables. -Easy to understand.	-Unable to handle linguistic terms as low, medium or high. -Dependent variable has to be binary/ dichotomous (0 and 1). -Unable to handle vague nature of binary observations. -Inability to handle ambiguity and vague relationship between independent and dependent variable.

(Abbod, et al., 2006;Dom et al., 2008; Gorzalczany, & Piasta, 1999)	Fuzzy-Neural	Cancer dataset Veterinary Medicine data set Human Medicine data set	-Needs transparent data. -Can model non-linearity. -Able to handle fuzzy, linguistic data.	-Cannot support high dimension problems. -Inadequate ability in explaining relationship between response and input variables.
(Dom et al., 2008; Hojati et al., 2005; Pourahmad et al., 2011; Mccauley-Bell et al., 1999)	Fuzzy-Regression	Cancer data set. Cumulative Trauma disorders disease data set SLE (Systematic Lupus Erythenalosis) dataset.	-Ability to perform optimal' input variable Selection. -Suitable for variables governed by vague and ambiguous relationship. -Easy to understand. -Adaptable to other nonlinear prediction problems. -Able to handle vague observations.-Scalable.	-Sensitive to outliers. -Multi-collinearity.

Table 2 Literature review on comparative analysis of data mining algorithms.

3.8 Comparison of Model table

For our research, comparison and evaluation of proposed model with existing models is vital. Therefore, studies are included with model evaluation methods. Model discrimination can be described with the help of c statistics with 95% confidence intervals. Comparison of various models according to literature is described in Table 3.

The performance of the model is assessed using overall performance measures, discrimination, and calibration. Traditional statistical approach is to quantify how close are predictions to actual outcome, using R-square and brier score. R-square is defined as the proportion of variation in the response variable that can be explained by predictors in the model (Gerds et al., 2008). Brier score is a quadratic scoring rule where the squared difference between actual and binary outcome is calculated. Overall performance of the model is quantified by measuring the distance between predicted and actual outcome. The distance between predicted and actual outcome is to quantify the overall model performance. These distance between predicted and actual outcome are related to the goodness-of-fit of the model. Better models have smaller distances between predicted and actual outcomes. Performance can be measured by discrimination and calibration. Discrimination measures how much the system can discriminate between the cases of readmitted patients”1”, and not-readmitted patients “0”. Discrimination can be assed using ROC curves. Calibration measures how close are the estimates to “probability of outcome”. It refers to agreement between predicted and actual outcomes. Recently, several new measures have proposed to assess the performance of the model. These are performance measure such as c-statistics for outcome of an event, which are refinements of discrimination measures. The c-statistics describes how well the model can rank “cases” and “non-cases” patients, but is not an actual function of predicted probabilities (Cook, 2007). Because this measure is solely based on ranks, it is less sensitive than measures based on the likelihood or other global measures of model fit. This characteristic makes it a poor choice for the selection of variables to be used in the predictive model. Using c-statistics for model selection could naively eliminate established risk factors from

risk prediction scores. As novel risk factors are identified, their dependence on c-statistics to evaluate their utility as risk predictors are not well-defined.

In our study, we have used model discrimination method for comparison and evaluation of proposed model with existing models. Therefore, we have reviewed studies with model evaluation methods in table 3. Different models with c-statistics are compared in this table. C-statistics as described above is equivalent to the area under receiver operating curve, is defined as proportion of times the model correctly discriminates a pair of high and low risk individuals. We also abstracted other operational characteristics such as sensitivity, specificity and predictive values for risk score cut-offs. A c-statistics of 0.50 indicates that the model performs no better than a chance, a c-statistics of 0.70-0.80 indicates modest or acceptable discriminative ability, and a c-statistics of greater than 0.80 indicates good discriminative ability (Kansagara et al., 2013). Selection of risk score cut off value is important in risk prediction models. It is rather difficult to define the optimal threshold. In ROC curves, we select a range of cut-offs for a sensitivity and specificity pairs. Risk threshold varies from (high-low) score, where low risk threshold may help in providing treatment to patients at risk of readmission at very early stage but may also lead to too many false positives. If the cost of treatment is not very high, then we can select low-risk threshold. On the other hand, if over treatment is quite harmful and expensive then we should use a higher cut-off before a treatment decision is made.

Author	Model Used	No of Patients	Re-admission within no of days	Rate of re-admission of patients	Range of Risk Scores	Model discrimination
(Billings et al., 2013)	Logistic Regression	576868	30 days	59.2% Positive Predictive Value	0-1	ROC Curve with value of 0.70
(Fialho et al., 2012)		26,655 of which 19,075 are adults where age>15		Sensitivity 0.68±0.02, specificity 0.73±0.03		AUC of 0.72±0.04, Rate of Risk of re-admission 4-11%
(Howell et al., 2009)	Case finding algorithm		12 months	Sensitivity 44.7%, Specificity 37.5%,	Risk threshold of 50.	Roc 0.65
(Billings et al., 2006)	Case Management	Age>65	12 months	Sensitivity 0.543,Specificity 0.722	Risk threshold of 50.	area under receiver operating curve 68.5%
(Philbin and Disalvo, 1999)		42,731, subgroup 9,112				P<0.10, Independent variables P<0.05
(Ottenbacher, et al., 2004)	Logistic Regression	9584 , with 51.6% females	3-6 months	Rate of risk re-admission 18.3%.		Area under ROC curve 0.68, Significant variables P<0.05,goodness of fit chi-square =11.32(df =8, P= 0.22)

Table 3 Comparison of Models for risk of re-admission.

3.9 Summary

In this chapter, other systematic reviews including data mining methods in health care such as artificial neural network, logistic regression, and fuzzy methods are described. As our proposed modelling approach uses fuzzy methods therefore detailed literature review of fuzzy methods is done. Literature on fuzzy methods includes study on fuzzy sets and membership function, fuzzy linear regression and fuzzy logistic regression methods. The risk factors in our existing models are also explained. We grouped studies based on list of predictors as given in various articles for patient's re-admission. Our study has limitations as although we used a clear and sensible search strategy, we may have missed some relevant publications. However, given the large number of studies included in our review, it is unlikely that overall conclusions would change meaningfully if any missed studies were included. The reviews of literature for this research has been built up based on the data mining and machine learning techniques for predictive modelling in healthcare resource utilization. Thus, based on this literature review we propose a framework which adapts fuzzy regression approach for predicting patients at high risk of re-admission. In chapter 4, we discuss a theoretical study for the proposed framework.

Chapter 4

4. Theoretical Study

4.1 Introduction

In this study, an exploratory research approach is employed to develop a framework for healthcare resource utilization. The framework is developed from synthesis of the literature reviews on existing prediction techniques specifically used for health care described in chapter 2&3. Guidelines given by the researchers on data mining and machine learning techniques in healthcare were the basis of the theoretical study and methodology in development of the framework. Research data includes large amount of imprecise observations. The delivery of healthcare services and quality of care depends on the availability of quality data. The goal of Poor knowledge about data, inaccurate and insufficient data can lead to increased costs, inefficiencies and poor financial performance. A clinical decision support system designed based on inaccurate or incomplete data, can give wrong clincal advice. These issues will have impact on payment, and may go much further than just finance. Furthermore, poor quality data inhibits clinical research, health information exchange, and quality measurement intiatives.

Problems will arise when there is an ambiguity of events or the degree to which an event occurs especially when the relationship between explanatory & response variable are vague (Shapiro, 2005; Dom et al., 2008). Fuzzy regression is useful when the available data is very limited or imprecise and when variables interact in uncertain manner (Vasant et al., 2002). Therefore, understanding of theoretical study behind the development of framework is important. Although knowledge of terms such as uncertainty, vagueness and imprecision is important for the proposed methodology, in our research we are using the term ‘uncertainty’ as the basis of our proposed framework.

In real world where theres is a substantial information, there is a great deal of uncertain or unknown information. Since health care system have some partially unknown parameters grey system theory is adapted to deal with such problems.

Because of the disturbance from both inside and outside and limitation of the current level of information grey system, rough set theory, fuzzy theories have gained popularity. The basic characteristic of uncetain system is incomplete or inaccurate information. A system which contains known values and uncertain unknown values is called a grey system (Zheng, 1993). Grey system theories became popular to deal with its ability to deal with the systems that have partially unknown parameters (Kayacan , et al., 2010). Rough sets theory studies uncertain system with accurate math methods (Liu & Sheng, 2012). The main idea of rough set is using the known knowledge base to describe and deal with the inaccurate and uncertain knowledge approximately. Compared with other theories, fuzzy theory is more suitable for human reasoning and natural language system. More detail explaination of fuzzy set theory is expalined in chapter 4.

This chapter describes the theoretical study for the framework. The component of this chapter includes study on concepts involved in proposing the framework. In the coming sections, elaboration of each component of the theoretical study is given. Later, three chapters (chapters 6, 7and 8) focus on the development of the framework, experiments and model validation based on this study. As data analysis plays an important role in the development of a framework, therefore detailed data analysis is given in chapter 5

4.2 Fuzzy methods

Fuzzy modelling techniques provide good concepts for dealing with uncertain information (Zadeh, 2005). A number of predictive models have been developed to predict patients at high risk of re-admission. Current predictive models use statistical regression techniques like classification & regression trees and logistic regression for predicting patients at risk of re-admission. In classical regression, parameters are assumed to be random variables with probability distribution function (Demir, 2014). Fuzzy regression is different from conventional regression techniques. Unlike statistical regression modelling that is based on probability theory, fuzzy regression is based on possibility theory and fuzzy set theory. (Beliakov, 1996; Zadeh, 2005). In fuzzy regression, the coefficients are subject to possibislitic approach, which tries to minimize the whole fuzziness of a model by

minimizing the total spreads of its fuzzy coefficients (Pourahmada, 2011). The fuzzy regression model is adapted when available data are uncertain or data available interacts in an uncertain manner. If this uncertainty is represented as randomness, this approach - combining randomness and fuzziness - leads to fuzzy randomness (Möller et al., 2002). The description of data uncertainty allows consideration of randomness and fuzziness together. (Shakouri et al., 2009) introduced a new approach based on non-equality possibility index, by which a minimum degree of acceptable uncertainty is found. (Hojati et al., 2005) reviewed the relevant articles on fuzzy regression and provided a new method for computation of fuzzy regression that is simple and gives good solutions. (Chen & Hsueh, 2009) developed an FRM model using the least-squares method based on the concept of distance. However, this method is sensitive to outliers. Therefore, (Yang & Liu, 2003) proposed new types of robust fuzzy least square algorithms (RFSLA) with a noise cluster for interactive fuzzy linear regression models. (D'Urso et al., 2013) proposed fuzzy linear regression model based on the Least-Median Squares –Weighted Least Square (LMS-WMS) estimated procedure. This procedure deals with data contaminated by outliers due to measurement errors. To handle the outlier problem, (Hung & Yang, 2006) proposed an approach to detect outliers. To handle the outlier problem, (Hojati et al., 2005) applies goal programming (GP) for estimating the linear regression parameters. (Hong et al., 2004) introduced the technique of regularization as a way of controlling the smoothness properties of the regression function. All the above mentioned approach cannot be used to deal with nonlinear problems. (Su et al., 2013) proposed non-linear regression model using Fuzzy Expectation Maximization (EM) algorithm based on maximum likelihood strategy.

4.2.1 Fuzzy logic and fuzzy set theories

The purpose of the study was to establish a roadmap which may help to forecast the future developments of fuzzy technology in healthcare (Abbod, et al., 2006). In recent literature a simple search for word "fuzzy" was used as a part of fuzzy sets or fuzzy logic (Abbod et al., 2006). Fuzzy logic theory and applications have vast literature. With regards to document literature, we can classify the development

of fuzzy theory and applications as having different phases: Phase 1 as concept of fuzzy theory as a tool for decision making, phase 2 as application of fuzzy theory in medical applications, and phase 3 where advances in fuzzy set theory and a few applications were developed (Vasant et al., 2002).The modern trend in medical applications problem deserves modelling of all relevant vague or fuzzy information involved in real decision making problems (Vasant, et al., 2002). Because of the inherent uncertainty in medical applications (Abbod, et al., 2007) developed an algorithmic solution. Currently, fuzzy technique is very much applied in the field of decision making. (Zadeh, 2005) referred to fuzzy set theory which was started by him in 1965. Fuzzy logic is based on fuzzy sets, linguistic variables, possibility distributions, and fuzzy rules (Shakouri G & Nadmi, 2009). Rules of healthcare include words like 'high risk' or 'severe pain' that are difficult to formalize and to measure. However, traditionally, mathematics uses crisp (well defined properties) i.e properties that are either true or false (Phuong & Kreinovich, 2001).When relationship among variables is complex, it cannot be always captured by traditional modelling techniques (Solomatine & Shrestha, 2009). Fuzzy logic based modelling approach has a significant potential to tackle the uncertainty problem and to model complex functional relationship (Lohani et al., 2006). Other advantages of fuzzy logic is its flexibility and tolerance to imprecise data (Zadeh, 2005)

4.2.2 Fuzzy sets and membership functions

Fuzzy set theory was first established by (Zadeh, 1965). He proposed the concept of fuzzy set which is useful in dealing with classes of problems where there is no sharp transition from membership to non-membership (Kim et al., 1996). Zadeh's fuzzy set theory, is a potential tool for dealing with uncertainty and imprecision. The characteristic of fuzzy sets is that the range of value of the membership relation is in the closed interval [0,1] of real numbers (Takeuti & Titani, 1984) .The theory of fuzzy sets provides a systematic framework for dealing with fuzzy quantifiers, such as many, few, most and linguistic variables like tall, small, high , low, old etc. A fuzzy set is a class of objects with a continuum of grades of membership (Zadeh, 2005).A fuzzy set is characterized by a membership function which assigns to each

object a grade of membership function between zero and one. The ordinary identification of membership function aims to identify the relation between underlying distribution and fuzzy data obtained from it.

Since the introduction of fuzzy set theory by Zadeh in 1965 several attempts to establish relationship between the grades of membership and the classical probabilities measures have been made (Beliakov, 1996; Zadeh, 2005).Hence, it is not so significant to discuss the detailed shape of membership function. Membership functions have the forms of triangle-shaped, bell shaped and trapezoid-shaped functions (Bellman et al.,, 1970; Tanaka et al.,, 1989). A membership function is a curve or shape that defines how each input space is mapped to a membership value (or degree of membership function) between 0 and 1. The curve or shape is known as a membership function and is often designate as μ. The output axis shows the transition from high to medium, and medium to low. The shape of membership function defines the transition from high to medium or medium to low. At the same time, patient can be at high or medium risk of readmission but to a different degree of certainty. Example is given in chapter 4.

There can be diverse shapes for membership functions for e.g trapezoidal, triangular, Gaussian membership functions etc. However, it is not so significant to discuss the detailed shape of membership function. Selection of the shape of membership function is based on the simplicity and easy to compute. Straight lines form the simplest membership function. Both triangular and trapezoidal membership functions are easy to compute. (Tamaki et al., 1998; Li et al., 2012) adopted the trapezoidal shape including triangular shape as the function of the membership function

We have chosen to adapt trapezoidal membership functions, which includes traingular membership function as the function shape of the membership function.

Definition1: Fuzzy sets as introduced by (Zadeh, 1965), are given by membership functions$\mu_A: X \to [0,1]$, where the value $\mu_A(x)$ indicates the degree to which the element $x \in X$ belongs to the fuzzy set A.The theory of fuzzy set provides a framework for dealing with linguistic variables such as high, medium or low.

In other words, a fuzzy set A in X is a set of ordered pairs

$$A = \{(x, \mu_A(x))\}, \quad x \in X \qquad (1.1)$$

Where $\mu_A(x)$ is the grade of membership of x in A and $\mu_A: X \to [0,1]$ is called the membership function.

A membership function is a curve that defines how each point in the input space is mapped to a membership value (or degree of membership) between 0 and 1. Our research uses triangular membership function and trapezoidal membership function which is shown in figure 3 and figure 4.

Definition2: A trapezoidal fuzzy number$A = (a_1, a_2, a_3, a_4)$ is said to be trapezoidal fuzzy number if its membership function is given by$\mu_A(x)$, where $a_1 \leq a_1 \leq a_3 \leq a_4$ where $a_1, a_2, a_3, a_4,$ are element of fuzzy numbers as shown in figure 3.

$$\mu_A(x) = \begin{cases} 0 & x \leq a_1 \text{ or } x > a_4 \\ \dfrac{x - a_1}{a_2 - a_1} & a_1 \leq x \leq a_2 \\ 1 & a_2 < x < a_3 \\ \dfrac{a_4 - x}{a_4 - a_3} & a_3 < x < a_4 \end{cases} \qquad (1.2)$$

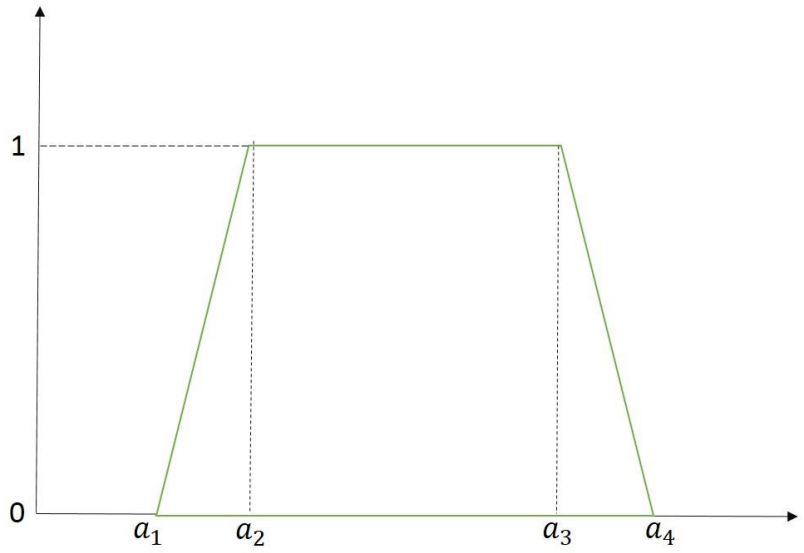

Figure 3 Trapezoidal membership function.

The interval a_2 to a_3 is known as the core. If a_2 is equal to a_3 , fuzzy number is referred to as "triangular" fuzzy number (TFN) which is defined in Definition 3. A fuzzy number is a number that has fuzzy properties, examples of which are the notions of "high", "relatively high", "low", "very low". The generalistic characteristic of a fuzzy number can be represented as shown in Definition 3.

Definition3: A triangular fuzzy number A is defined as a triplet (a_1, a_2, a_3) membership function $\mu_A(x)$ can be defined as

$$\mu_A(x) = \begin{cases} \dfrac{x - a_1}{a_2 - a_1}, a_1 \le x \le a_2 \\ \dfrac{x - a_3}{a_2 - a_3}, a_2 \le x \le a_3 \\ 0 \quad , \ otherwise, \end{cases} \qquad (1.3)$$

Where$a_1 \le a_2 \le a_3$; the elements of the fuzzy numbers are real numbers and its membership function $\mu_A(x)$ is the regular function, showing that the membership degree to the fuzzy set, a_2 represents the value for which $\mu_A(a_2) = 1$, and a_1 and a_3 are the most extreme values on the left and right of the fuzzy number$A$ (Li, et al., 2012) as shown in figure 4

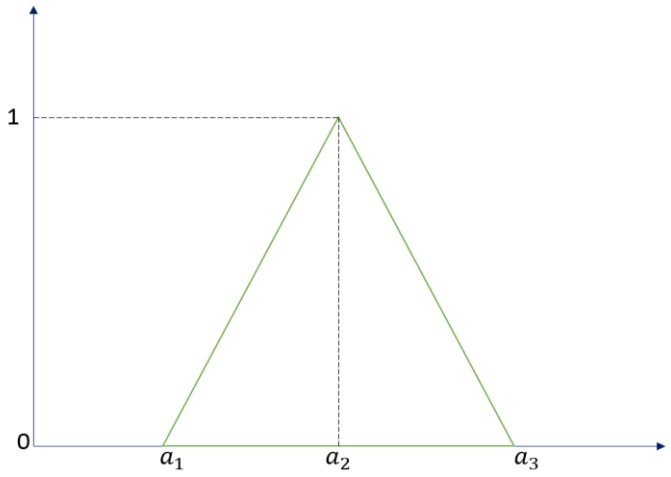

Figure 4 Triangular membership function.

For the proposed, fuzzy linear regression model the membership functions are determined by three points namely the centre point, left end and right end point. Figure 4 shows the membership function of the data with membership function, centre$(a_2)$and two end points as (a_1) and(a_3). As indicated, the salient features of triangular fuzzy number (TFN) are its centre, its left and right spread. When the two spreads are equal, the TFN is known as symmetrical TFN (STFN).

Definition4. An interval-valued triangular fuzzy number is a fuzzy intervalA^L, A^U, shown in figure 7, where both the lower-bound $A^L = (a_1^L, a_2^L, a_3^L)$ and the upper bound $A^U = a_1^U, a_2^U, a_3^U$are triangular fuzzy numbers and $a_1^L \leq a_2^U$ (Li, et al., 2012; Wei & Chen, 2009) which is shown in figure 5.

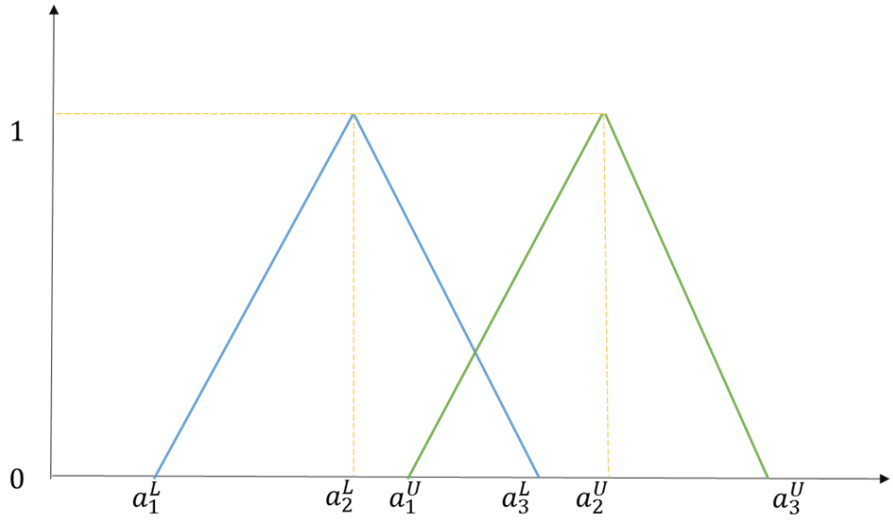

Figure 5 Interval valued triangular membership function.

Definition5: An interval valued trapezoidal fuzzy number as shown in figure 6 can be represented by,

$A = [A^L, A^U] = [(a_1^L, a_2^L, a_3^L, a_4^L, w_A^L), (a_1^U, a_2^U, a_3^U, a_4^U, w_A^U)]$ where $(a_1^L \leq a_2^L \leq a_3^L \leq a_4^L), a_1^U \leq a_2^U \leq a_3^U \leq a_4^U$, A^L denotes the lower IVFN and A^U denotes the upper IVFN. (Li, et al., 2012), Where $0 \leq w_A^L \leq w_A^U \leq 1$ and $0 \leq w_B^L \leq w_B^U \leq 1$

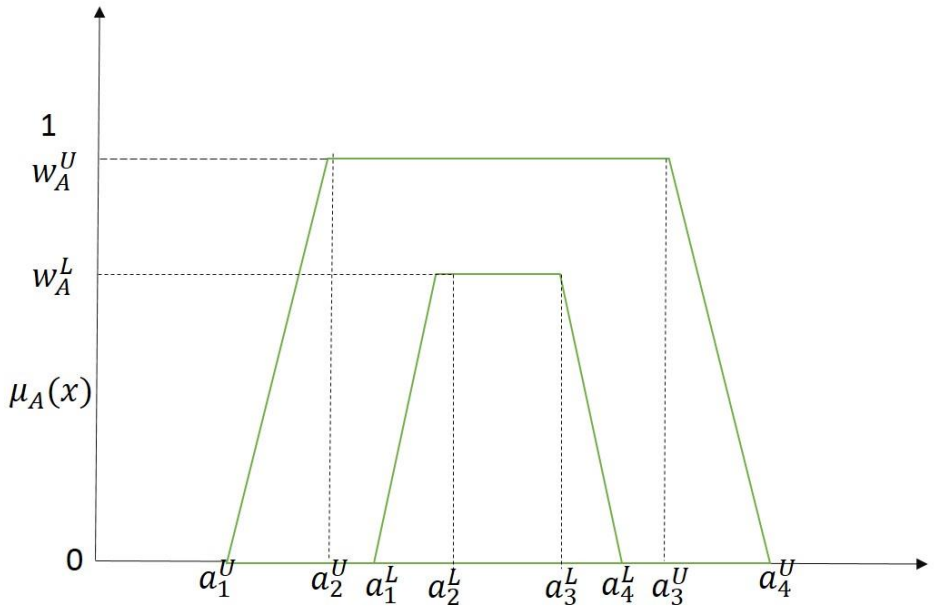

Figure 6 Interval valued trapezoidal membership function

.

4.3 Preliminary Theory

In statistics, linear regression analysis is a powerful method for studying the linear relationship between one response variable $\hat{Y}$ (dependent variable or output variable) and a set of explanatory variables $X_1,,,,,,,,X_K$ (independent variables or input variables) (D'Urso , et al., 2013).Classical statistical regression has many applications, problems may occur in some situations. Linear Regression method is extremely sensitive to outliers. Other situations are:

- Imprecise Information
- Uncertain data
- Vagueness in the relationship between input and output variables.

In recent years, there is a growing literature that formalizes the linear regression model in fuzzy domain, in which model parameters and/or data are fuzzy, or imprecise or vague (D'Urso , et al., 2013). Abundance of vague observations in healthcare studies, motivate us to think about a proper model in a fuzzy environment. These are the situations fuzzy regression was meant to address.

4.3.1 Approaches of Fuzzy Regression analysis

Fuzzy regression has been proposed to evaluate the complex relationship between dependent and independent variables in a fuzzy environment. Fuzzy modelling techniques provide good concepts for dealing with uncertain information (Bisserier et al., 2010). Fuzzy regression is different from conventional regression technique in the sense that it is a non-statistical method and is based on possibility theory. This methodology is appropriate for dealing with uncertain and vague information in systems. This section provides an introduction to fuzzy linear regression. The topics include motivation, the components of Fuzzy Regression (FR), fuzzy numbers, membership functions, and fuzzy output. In order to adapt fuzzy regression method in our proposed methodology, it is vital to understand preliminary theory concepts which are as follows:

There are two approaches to fit the fuzzy regression analysis.

Possibilistic Model: The first one is possibilistic approach introduced by Tanaka. According to this approach, fuzzy regression coefficients are estimated by minimizing the total spread of its fuzzy coefficients, subject to including the data points of each sample within a specified data interval (D'Urso et al., 2013; Shapiro, 2005).

The second approach is the Least Square (LS) approach which minimizes the distance between the output of the model and the observed output, based on their modes and spreads. The estimation procedure consists of finding the linear model which best approximates the observed data in a given space, taking into account the fuzziness of data (Chen & Hsueh, 2009; Pourahmada, 2011; Arslan, 2011).

There are three ways to develop a fuzzy regression model:

Case1: Independent variables(x) are numbers (=crisp), and response variable(y) is fuzzy.

Case 2: Independent variables(x) are fuzzy, and response variable(y) is also fuzzy.

Case3: Models where the relationship of the variables is fuzzy.

4.4 Fuzzy Linear Regression methods

(Tanaka, 1989)first proposed a study of fuzzy linear regression (FLR) model. Indeed, unlike statistical regression based on probability theory, fuzzy regression is based on possibility theory and fuzzy set theory. By classical statistical technique, the observations of either the response variable or the explanatory variables follow certain probability distributions (Beliakov, 1996). Fuzzy regression can be classified into two distinct areas, the first proposed by (Tanaka, 1989) minimizes the total spread of the output, is called possibilistic Regression. The second approach, is proposed by (Diamond, 1988), minimizes the total square error of the output and is called the Fuzzy Least Square method.

The advantage of Tanaka's possibilistic model is in its simplicity in programming and computation (Tanaka, 1989)while Fuzzy Least Square Estimation [FLSM] is good for minimizing errors between the given observed and estimated values (Tanaka, 1989). The possibilistic regression analysis uses a fuzzy linear system as a regression model whereby the total estimated vagueness of the estimated values of the dependent variables is minimized (Pourahmada, 2011) (Tanaka, 1989).

Case1: Independent variables(x) are numbers (=crisp), and response variable(y) is fuzzy.

The linear regression model is the most frequently used form in regression analysis for expressing the relationship between one or more explanatory variables and response. Fuzzy linear regression analysis was proposed by (Tanaka, 1989) to determine the fuzzy linear relationship:

$$\hat{Y} = A_0X_0 + A_1X_1 + A_2X_2 + \cdots + A_KX_K \qquad (1.4)$$

Where each regression coefficient A_j , $j = 0, \ldots\ldots, k$, was assumed to be symmetric triangular fuzzy number with centre α_j (having membership =1) and half width a_j , $a_j \geq 0$. This can be shown with the help of figure 7

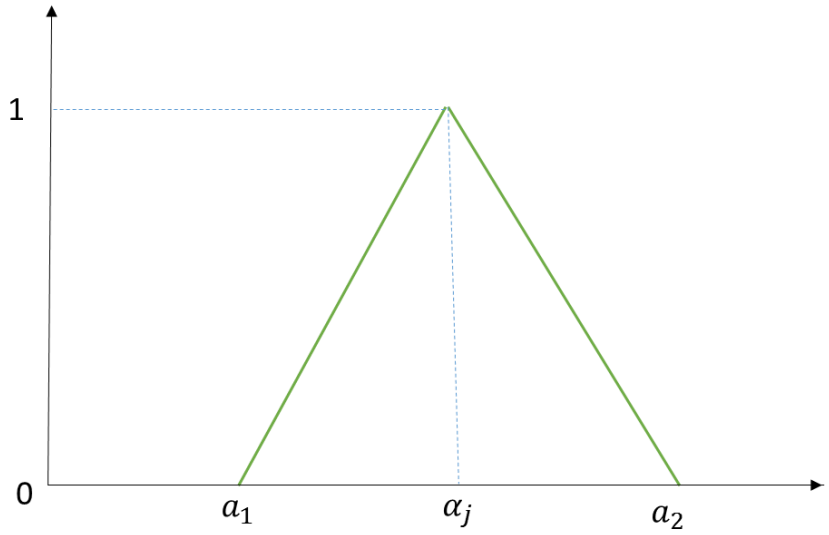

Figure 7 Triangular membership function with centre α_j

To model binary relationship between a binary response variable and a set of explanatory variables, let the input –output data consist of observations

$$(x_{i0}, x_{i1}, \ldots . x_{ik}, \widehat{Y}_i), 1 \leq i \leq m, \qquad (1.5)$$

Where x_{ij} $j = 0,1, \ldots ., k$ are real crisp values in R and $\widehat{Y}_i$ is a fuzzy observation detecting the status of each case relative to binary response categories i.e it takes two labels: approximately 1 or approximately 0 instead of 1 or 0.

4.4.1. Example of application of fuzzy methods in health care.

(Nagar & Srivastava, 2008)proposed an adaptive technique in the prediction of dichotomous response variable by combining fuzzy concept with statistics logistic regression. (Dom et al., 2008)developed a learning system for the prediction of dichotomous response variable by combining fuzzy concept with classical regression technique. In this model, fuzzy linear regression and logistic regression theories are combined to produce an adaptive fuzzy regression model. Their study was applied to fuzzy independent variables of risk factors in multiple logistic regression model (MLRM). The purpose of the study done by (Dom et al., 2008) is to present the use of fuzzy regression models in the prediction of oral cancer susceptibility as a function of demographic profiles and risk habits. (Ozdamar et al., 2005) developed a multivariate fuzzy linear regression (FLR) model for predicting

aggregate annual LOS. (Dom et al., 2008) adapted fuzzy regression method to predict the length of survival of AIDS patients based on their CD4, CD8 and viral load counts. As people have their interests in their health recently, fuzzy regression application in medical health domain has been one of the most active research areas. (Kunjunninair, 2012) developed a weighted fuzzy rule based clinical decision support system for the diagnosis of heart disease. (Tsipouras et al., 2008) and Setiawan et.al (2009) have developed a fuzzy decision support system for diagnosis of coronary artery disease. (Ponzo et al., 2016) uses fuzzy regression discontinuity designs to evaluate the impact of cost sharing on the use of health services. (Sener et al., 2011) describes fuzzy multi-objective programming approach (Setiawan, 2014) that incorporates imprecise information in QFD (Quality Function Deployment) tool to determine the level of fulfillment of design requirements.

(Abbod et al., 2007) studied Neuro-fuzzy modelling system (NFM) in health care. The study shows that lack of transparency of artificial neural networks can be overcome by a neuro-fuzzy modelling system. The study also proves that Neuro-Fuzzy model outperforms other models such as Bayesian Belief Network and conventional statistical methods in its accuracy.

4.4.1 Advantages of fuzzy regression methods

- Fuzzy regression is useful in estimating the relationship between independent and dependent variables when the available data is uncertain and imprecise. (Shapiro, 2005).
- The fuzzy linear regression is an alternative to statistical linear regression in estimating regression parameters when statistical model is with poor model specification due to uncertain relationship among variables (Ozdamar et al., 2005).
- Fuzzy regression gives better performance in case of imprecise and uncertain data (Beliakov, 1996; Möller et al., 2002)

4.4.2 Limitations of fuzzy regression methods

Some of the limitations of fuzzy regression models are:

- The fuzzy linear regression may tend to become multi collinear as more independent variables are collected (Shapiro, 2005; Kim, et al., 1996).
- The original fuzzy regression model was extremely sensitive to outliers (D'Urso , et al., 2013; Shapiro, 2005; Peters, 1994)
- There is no proper interpretation of the fuzzy regression interval (Shapiro, 2005).

4.5 Least square estimation method

The fuzzy least squares approach, which is proposed by (Diamond, 1988), is an extension of the ordinary least squares based on a new defined distance on the space of fuzzy numbers. Diamond (1988) was the first to implement fuzzy least square regression (FLSR) using distance measures and his methodology is the most commonly used. Given triangular fuzzy numbers (TFNs), it provides a measure of the distance between two fuzzy numbers based on their centre, left spread and right spread. The definition of a function which describes well the distance between two fuzzy numbers is somehow difficult (Pourahmada, 2011). An obvious way to bring fuzzy regression (FR) more in line with statistical regression is to model fuzzy least square regression (FLSR) along the same lines. The linear least-squares technique for m pairs of real crisp numbers $(x_i, y_i), 1 \leq i \leq m$, consists of finding $a, b \in R$ such that the sum

$$r(a,b) = \sum (a + bx_i - y_i)^2 \qquad (1.6)$$

will be minimized. In the case of fuzzy data, $u = A_1 + A_2 v,\ u, v \in E$, we seek numbers A_1 and A_2 such that the distance between observations and estimations will be minimized.

4.6 Interval-valued Fuzzy numbers

In real life, a person may assume that an object belongs to a set of certain degree, but there may be uncertainty about the membership degree of an object belonging to a set. When something is uncertain, such as measurement, use of fuzzy sets which represents uncertainty by numbers in the range [0,1] makes more sense than conventional sets (Chen & Lai, 2011; Bellman & Zadeh, 1970) and (Zadeh, 1965)

were the first to introduce the theory of fuzzy sets in problems of decision making as an effective approach to treat vagueness, lack of knowledge, and ambiguity inherent in decision making process (Li, et al., 2012; Zadeh, 2005). (Zadeh, 1965) and (Yager, 1986) extended the concept of fuzzy sets and adopted interval valued fuzzy numbers (IVFNs) for handling uncertainties arising from incomplete, vague or imprecise information. Because uncertain is an attribute of information, it appears to be a more applicable method for health care systems to handle such health system variables by using IVFNs. In fuzzy set theory, it is often difficult to identify any opinion as a number in the interval [0, 1]. Therefore, to represent the degree of certainty of opinions by interval value fuzzy numbers is more appropriate.

4.6.1 Variable selection and multi-collinearity

Multi-collinearity is the situation in which two or more input variables in a multiple regression model are highly correlated. (Coppi, 2008) extended the robust fuzzy regression model to deal with multi-collinearity problem in input variables. Variable selection is always a focus of much research in machine learning tasks including classification and regression prediction. The term 'features' refers to the attributes, properties and characteristics of input variables. An appropriate variable selection enhances the effectiveness and domain interpretability of a prediction model. In traditional regression framework, several procedures have been suggested for choosing suitably the set of explanatory variables. The most common methods include forward selection technique, backward elimination technique and stepwise procedure (D'Urso et al., 2013).

In fuzzy regression, a model could be established by using more than one independent variable. However, large number of variables could lead to problems like correlation among variables which makes the fuzzy regression model multi-collinear. The idea is to include as many independent variables as possible but at the same time avoiding co-linearity problem in input data variables.

In our research we have conducted experiment to show the multi-collinearity problem among variables. In chapter 6, detailed algorithm for multi-collinearity problem is discussed. In our algorithm, an Interval Valued Fuzzy Number approach is proposed for dealing with multi-collinearity problem in health system variables.

4.7 Summary

The purpose of this thesis is to propose a framework which adapts fuzzy regression methods to predict patients at risk of re-admission. In order to develop a framework, theoretical study on fuzzy regression methodology plays an important role. In this chapter, we have described, fuzzy regression method with its limitations and advantages. Fuzzy regression method estimates the uncertain relationship among dependent and independent variables. This method deals with uncertain and imprecise data. Therefore, understanding of concepts such as vagueness, imprecision and uncertainty is also essential. In order to develop a methodology, dealing with uncertain data is also required. Variable selection is always of focus in development of a model. Significant independent variables act as input for a predictive model. Therefore, data analysis for significant input variables and dependent variable is also vital for any model. In the next chapter, detailed analysis of data is done.

Chapter 5

5. Data analysis and Preparation

5.1 Introduction

This chapter contains a brief outline of how the data was prepared and manipulated for our research. Data analysis was based on the algorithm called PARR (patients at risk of re-hospitalisation) which identified patients who are likely to be at risk and therefore require interventions. The algorithm used data from 1999/2000 to 2005/2006 of England hospital inpatient episode statistics data. A random sample of patients with an emergency inpatient triggering admission in 2004/2005 were analysed of having a re-admission in the following twelve months (i.e. upto end of 2005/2006) by the PARR algorithm. Selection of variables for our proposed approach was similar to the variables obtained from PARR1 or PARR2. Significant variables for re-admission in the next 12 months were obtained by analysing the patient's previous years of hospitalisation data from 1999/2000 to 2003/2004 prior to the triggering of re-admission in 2004/2005.

5.2 Data Preparation

This section contains a brief outline of data preparation and manipulation for the research work. Data is extracted from MySQL workbench using SQL queries from the Hospital Episode Statistics (HES) database. Each record in the HES database is episode based, which represents period of time a patient is under care of a particular consultant. A patient may have more than one episode in a spell. Many spells finish with this episode, but if the patient moves to the care of another consultant, a new episode begins. However, admission date for each new episode in a spell will remain be equal to the episode start date of the first episode within that spell. Episode numbers increase by 1 for each new episode until the patient is discharged. If the same patient returns for a different spell in hospital, epiorder is again set to 01. Admissions are calculated by counting the number of times epiorder is 01.

This chapter describes the process of extraction and manipulation of 109,243 admission records. Data is extracted using MySQL workbench and imported in SPSS for manipulation and preparation. Input variables are recoded to obtain independent and dependent variables. Statistical analyses were carried out on this dataset to derive significant independent variables from which we could predict a subsequent re-admission. Predictive models are generally built on a data set consisting of dependent variables (re-admission) and a range of independent variables from records of patient in previous years. The dependent variable is fuzzified using the triangular or trapezoidal membership function. Fuzzy variable "risk of readmission" is derived after fuzzifying response variable "re-admission". This derived fuzzy variable is a response variable for our proposed model. Risk of readmission can have a range of values from high, medium to low risk of readmission. For performing statistical analysis, the dataset was divided into training set and testing dataset. This chapter focuses on the processes that were carried out before the application of the predictive modelling techniques.

5.3 HES Data

Each of the HES year data tables for financial years (April to March) 1999/2000 to 2005/2006 contained millions of records representing individual episodes of inpatients in England during these years. The data had to be prepared and analysed for future analysis. MySQL queries were used to extract a sample of 100,000 emergency admissions that started and ended in 01/04/2004 and 31/03/2005. The same software was used to extract the next emergency admission which was within 1, 6 and 12 months of the discharge date of the triggering re-admission for these patients. MYSQL was used to extract prior hospitalization data for these patients in the five years leading up to their triggering re-admission date. All of the variables defined in this section were derived for later use in predictive modelling. This means data for about 100,000 patients was in one file with all required variables.

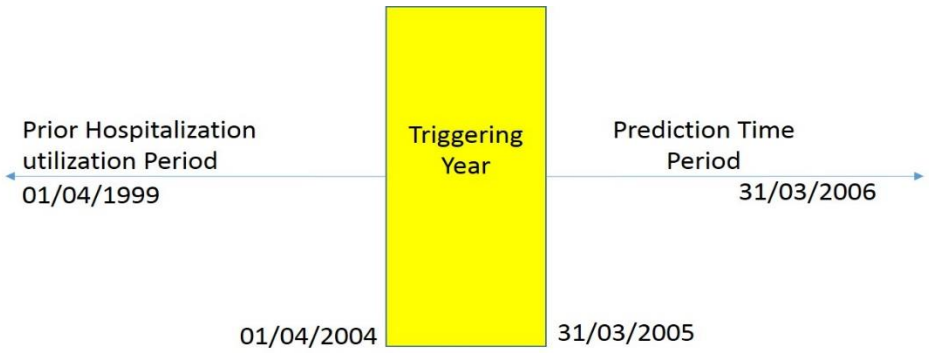

Figure 8 Time frame for the HES data in algorithms

5.3.1 Data Sample

Approximately 3.5 million records of emergency inpatient admissions that started and ended in 2004/2005 within England were extracted. A random sample was selected from total population of 3.5 million records. SQL query A was used to extract a sample of 3.5 records of emergency admissions that started and ended in 2004/2005. **SQL exhibit A** returns data for the last episode (given by discharge methods 'dismeth' other than 4 (died), 8 (not applicable: patient still in hospital) and 9 (not known: a validation error)) of all emergency inpatient admissions (given by admission methods 'admimeth' of 21, 22, 23, 24 or 28 which all stand for emergencies) that started and ended between 01/04/2004 and 31/03/2005. This includes patients who have a valid sex code of 1(male) and 2 (female), valid date of birth, and valid discharge date. Data excluded were patients with invalid date of birth or invalid triggering admission date, or invalid discharge date.

However, we only wanted to work with a smaller sample of these records to predict readmission so **SQL exhibit B** was used to take a random sample of records. The episode key 'epikey' is an eight-digit number which identifies a patient's episode of care and the random sample of 109,243 were selected by choosing all of those records from the 3.5 million whose episode key ended in 01, 31 or 61, and with valid discharge dates. After running SQL exhibit B on a sample of 3.5 million records (extracted after SQL exhibit A), we received a sample of 109,243 records.

Epikey is the unique record identifier that is created by HES system. The digit stores a decimal number of 8 or 9 digits but can be upto 14 digits. We need a sample of

record from overall records to carry out the analyses. To select a sample of records, I have randomly chosen records which ended in 01, 31 and 61. These codes (01, 31,61) are just last two digits of unique record identifier and they are chosen randomly to give a sample of records. After selection, sample of record returned is 109,504. This sample is slightly different from 109,243. This is due to the fact that further 261 records were removed, as these records did not have valid discharge dates.

The final sample size is of 109,243 records as patients with invalid discharge dates and missing values are removed from the final sample. Missing values are treated using is.na function in R. Missing values and system missing values are interpreted as NA. In R, missing values are treated using is.na function. In our analysis, missing values are ignored using is.na functionThe extracted data for the 109,243 included the variables which gave us the information like the age at start of admission, gender, triggering admission date, discharge date and diagnostic conditions.

This includes patients who

- had a valid sex code of 1(male) or 2(female)
- had a valid date of birth
- had a discharge method of other than 4 (patient died), 8 (not applicable: patient still in hospital) or 9 (Not known: a validation error) for their triggering admission
- had a valid discharge date.

5.4 Data Preparation and Manipulation

A random sample of 109,243 of these records is used to create the variables in this section which are used in the model from which we could predict a subsequent re-admission.

5.4.1 Variables used in the Analysis

All of the variables were derived from HES data to predict re-admission of sample of patients taken from 2004/2005.

5.4.2 Independent variables

Independent variables (such as age, gender, ethnic origin etc.) are created and used to predict the dependent variable (i.e. whether the patient has a re-admission). The following independent variables were created by looking at the data from triggering admission.

In stage 1, binary variables were derived for gender and age at admission. Binary variables are added to the file with the sample of 109,243 records. The variables in HES data were re-coded by using "recode" function. The variable "sex" was denoted in HES data as males by 1 and females by 2. This was transformed into variable sex_recoded by using recode into different variable so that males are recoded as 1 and female were recoded as 0.

The variable 'start age' in the HES dataset gives us the age (in years) at the start of the current episode. Patient's age at the time of triggering admission was split into groups and binary variables were created to show which group a patient was in. Binary variables for age on admission were derived to indicate whether patient was young, not so young, less old or old. The start age variable was used to determine the age at start of the triggering admission.

The following binary variables were created to group the patients.

5.4.3 Patient characteristic and demographic variables

- Gender of the patient
    - Sex (1=Male, 0=Female)

- Age of the patient

Age bands are defined so that the patients could be divided into age groups for creating and manipulation of age variable. Age groups are defined for creating binary variables for age. Input variables are coded as binary variable for implementing various algorithms (Fuzzy regression, logistic regression, decision tree and Neural network).

Start age in HES data dictionary defines age of a patient at start of episode. For patients under 1-year-old, special codes in the range 7001 to 7007 apply. Patients

under age of one year are considered as very young and patients with age greater than 1 year are young patients. For our analysis, we have considered patients within age band of 0-18 as young but further they can be classified as very young and young. In our analysis, we are concerned only for young patients. Patients could be classified as young, and not so young. As in our research, we are not fuzzifying input variables therefore, we have not considered young and not so young bands.

Four binary variables for age on admission were created to indicate whether the patient fell into one of the age groups listed above. For example, those who were aged 16 on admission had the binary variables aged_0_17 set to 1 and the other four binary variables for the age groups set to 0.

- Age 0 to 17 (1 = yes, 0 = no)
- Age 18 to 39 (1 = yes, 0 = no)
- Age 40 to 64 (1 = yes, 0 = no)
- Age 65 to 74 (1 = yes, 0 = no)
- Age 75 plus (1=yes,0=no)

- Ethnic origin of the patient

The variable ethnos in the HES dataset gives us the ethnic origin code for the patient. There are many different codes used to represent different ethnic origin groups. However, patients that were white were coded into the variable of white = 1 and those who were non-white were coded as white =0.

- White (1=yes, 0=no)
- Black (1=yes, 0=no)
- Mixed (1=yes, 0=no)
- Asian (1=yes, 0=no)
- Other or unknown (1=yes, 0=no)

5.4.4 Deriving the dependent variable

Stage 2 involves joining the results from stage 1 to the tables of data for 2004/2005 and 2005/2006 to extract the date of the next emergency admission within 1

month, 6 months (180 days) and 12 months (365 days) of the discharge date of the triggering admission. Each of the109, 423 records were recoded into binary variables called Re-admission_30, Re-admission_6, and Re-admission_12. These variables were recoded for whether or not they had a re-admission within 30 days, 6 months or 12 months. Those records that did have a re-admission within 180 days were recoded as 1 and those that did not have a re-admission the variable was set to 0. Similarly, binary variables were also coded if the patient had re-admission within 12 months.

The dependent variable created are:

- Re-admission within 12 months (1=yes, 0=No)

The following 2 binary variables were also created as we wish to look at re-admission within different time frames (specifically 30 days, 6 and 12 months). Additionally, these two variables are created to test our algorithm. But our main focus is on readmission within 12 months.

- Re-admission within 30 days (1=yes,0=no)
- Re-admission within 6 months (1=yes,0=no)

5.4.5 Deriving the remaining independent variables

State 3 involved joining the results achieved after stage 2 to the data for years 1999/2000-2004/2005 to extract data on the previous episodes for in-patient admissions in the five years prior to their triggering admission date.

Patients with chronic and most common diseases were considered for our analysis. The HES data contains 14 diagnosis fields labelled as diag_nn for each of the episodes. The value of nn will range from 1 to 14, and not all 14 fields have values. Variables were created for each of the conditions to record whether each of the inpatient admissions had the condition in the current admission or in previous five years. Each of the diagnostic variables has to be scanned for all admissions in the five-year time to check for the conditions.

Each of the binary variables was coded as 1 if the patient had the condition or 0 otherwise.

- Alcohol abuse
- Anaemia
- Angina
- Atrial fibrillation
- Cancer
- Cerebrovascular disease (CVD)
- Congenital disability
- Congestive Heart failure (CHF)
- Connective tissue disease/rheumatoid arthritis(CTDRA)
- Chronic obstructive pulmonary disease(COPD)
- Development disabilities
- Diabetes
- Drug abuse
- HIV/AIDS
- Hypertension
- Injury from fall
- Ischaemic heart disease(PVD)
- Renal failure
- Respiratory function
- Sickle cell anaemia

Conditions are identified in the HES data by ICD 10 codes. ICD 10 stands for International Classification of Diseases Tenth Edition. This method of condition classification is standard for recoding health problems and conditions in health data. Table A3.1 in Appendix 3 contains a list of the ICD 10 codes for the conditions used in this thesis.

There is one variable (severity index) related to this section which calculates the Charlson Comorbidity Severity Index(CCSI) for each triggering admission as mentioned in Appendix 3. The CCSI is widely used to indicate how severe conditions

are. The comorbidity severity index score allocates severity scores to diseases. Charlson score designed for breast cancer survival was used, as it gives modest benefits overall. Weights assigned to each comorbid conditions are more stable in Charlson score designed for breast cancer. Modified Charlson coded and adjusted weights may give better fit and discrimination, but its weights are less consistent across all patient groupsFor each of the 109,243 rows of data it was recorded whether the patients had any of the conditions at any time during previous five years (Table A3.2 of Appendix 3). This was incorporated into this analysis by summing up the total severity of conditions that the patient had in the triggering admission and in the 5 years prior to that point.

Severity index score is calculated by summing up all conditions and multiplying by weighing factor to produce the total severity score (Calculation method is shown in Appendix 3 with conditions given in Table A3.1). For example, if the patient was flagged as having congestive heart failure (which carries a severity weight of 1) and HIV (which carries a severity weight of 6) only during their last five years then they would have a total severity weight of 7. Patients with higher total severity index scores either had more severe conditions or just had multiple conditions.

Another variable which was included was whether the patient had a current or prior emergency admission for a reference condition (1 = yes, 0 = no) in the previous five years. The reference conditions are those which are thought of as being more likely to result in re-admissions and are defined using the Healthcare Resource Group (HRG) codes within Table A3.3 in Appendix 3.

5.4.6 Patients Prior Hospital Utilisation

In our research, we are considering prior hospital utilisation for patients in the last 5 years prior to the triggering admission date. Dependent variable is calculated for patient's readmission in the past 30 days, 6 months or 12 months, as this has helped us in giving readmissions in different periods. Readmission within 30- days helps us in understanding the chances of readmission within short time after discharge, and if it could be avoided by providing better interventions. To evaluate patients at risk of readmission within 30, 6 or 12 months, we evaluate independent variables based

on past 5 year's historical data. The following numerical variables were based on the prior hospital utilisation for the patients in the last five years prior to the triggering admission date in 2004/2005.

- Number of re-admission in the previous 30 days.
- Number of re-admissions in the previous 6 months.
- Number of re-admission in the previous 12 months.
- Average number of episodes per spell for patient's re-admission.

5.5 Statistical Analyses

Differences between patients readmitted and those not readmitted were analysed using chi-square univariate tests. Patient dataset with independent variables and dependent variable for all years from 1999/2000 to 2004/2005 was joined with approximately one million records. The relevant variables were based on a broad range of measures used in the novel algorithm. These included number of admissions to the hospital according to a time interval prior to current admission (30,180 or 365 days), number of episodes per spell in the prior admissions, a range of diagnostic categories and diagnostic groups. The reduced number of variables ultimately included in the algorithm is based on the significant predictors based on statistical analyses. Predictive models are generally 'trained' on a data set consisting of dependent variables (Readmitted patients in hospitals) and a range of independent variables from record of patient in previous years. Among the patients in the sample dataset, univariate analyses was carried out to determine which patient characteristics and health outcomes had significant impact for hospital re-admission. Significance of variables responsible for risk of readmission was evaluated using statistical tests for e.g pearson correlation coefficient. Logistic regression was considered to calculate the probability of an event given risk factors. We also carried out univariate analysis (chi-square) to identify variables significantly associated with readmission. All variables with value of $p<0.05$ were included in our multiple regression model. All patients with significant or borderline statistical relationship with rehospitalisation at the univariate level ($p \leq 0.05$) were

entered as independent variables in the predictive model for re-admission. The p value ≤ 0.05 was considered statistically significant.

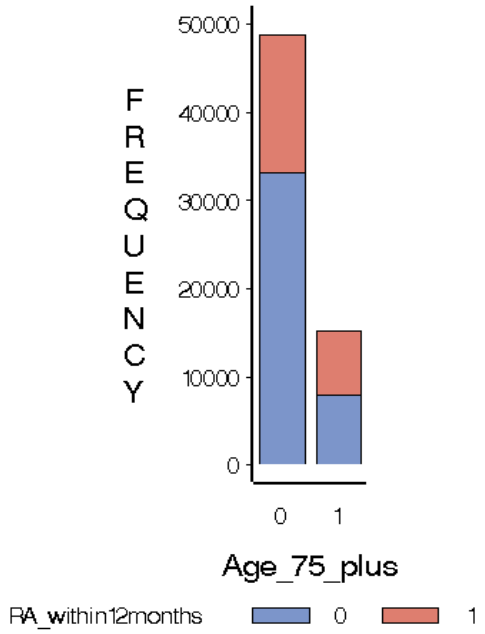

Figure 9 Age 75+ at admission and re-admission within 12 months

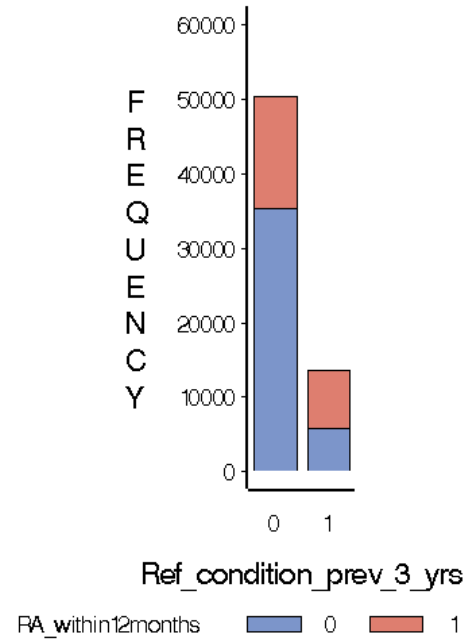

Figure 10 Presence of a reference condition and re-admission within 12 months

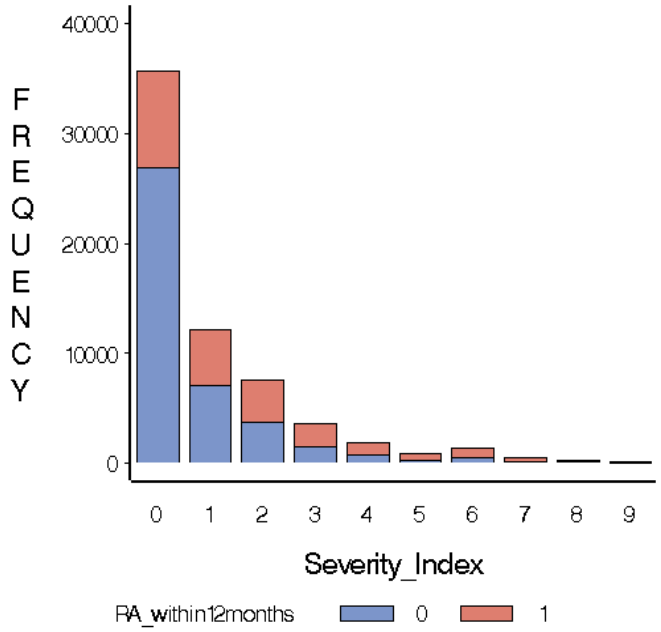

Figure 11 Total severity index score and re-admission within 12 months

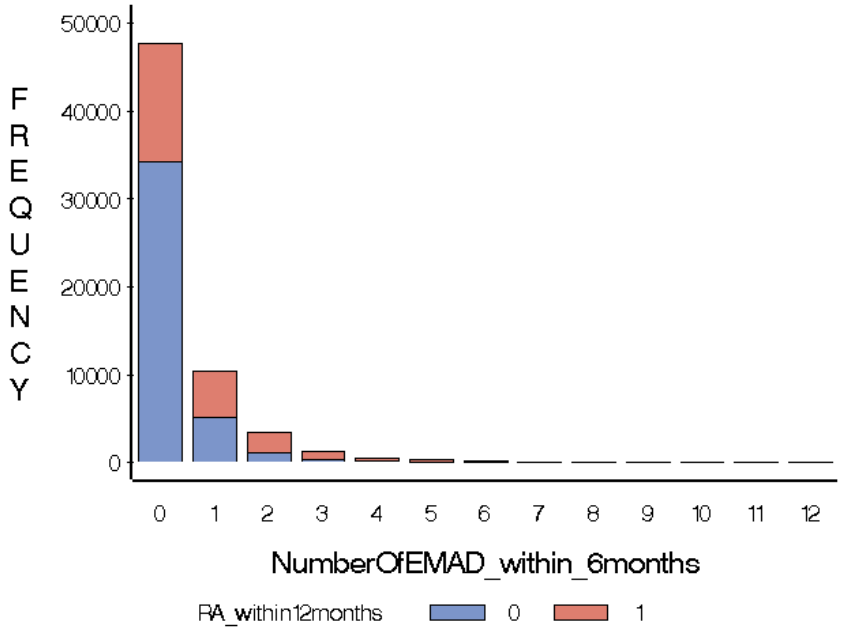

Figure 12 Number of emergency admissions

A figure 12 show that patients are more likely to have a re-admission if they are aged 75 or over in their triggering admission as re-admission (the pink shaded area) is more prevalent in the aged 75 or over group. Figures 10, 11 and 12 suggest that patients are more likely to have a re-admission if they have a reference condition, a high severity index total score or more admissions in the previous 12 months.

5.6 Data Quality

Amongst the sample of 109,243 patients with re-admissions in 2004/2005 there were no duplicate records for the same patient having re-admission starting on the same date. All of the duplicate records for the unique HES Id were removed from the dataset. There were also no missing (also known as null) values in the data fields. When recording historical information such as the number of previous admissions, duplicate entries for the same admission were ignored to ensure that the admission was counted only once.

5.6.1 Data Pre-processing

The dataset was partitioned into two sections as all predictive models are constructed on a training dataset and validated or tested for performance on a validation dataset. Therefore, the full dataset of 109,243 rows was randomly split using a random selection so that 60% of the rows (65,547) were used for the training dataset and the remaining 40% of rows (43,698) were used for the validation dataset.

5.6.2 Removing outliers from the dataset

Unusual or extreme observations (outliers) for interval or continuous (non-binary) variables are usually removed from training datasets prior to the application of predictive algorithms to ensure that the models are built using stable and consistent data. Therefore, the extreme top and bottom 0.1% of values for the interval variables were removed from the training dataset.

There are more statistical approaches for detecting outliers. One such single dimensional method is Grubb's method which calculates a Z value as the difference between the mean value and query value divided by the standard deviation for the attribute. Mean value and standard deviation are calculated from all attribute values including the query value. One of the simplest outlier detection techniques used is box plot to prinpoint outliers in univariate or multi-variate datasets. Other oulier detection methods could be proximity-based techniques, which are based on distance-based measures such as Euclidean distance and Mahalanobis distance. P-Plots can also be used to detect outliers in the dataset.Outliers are not removed

from the validation dataset as each of the models should be tested on actual data to determine their true performance.

5.6.3 Selecting the important independent variables to predict re-admission

Before running the predictive modelling algorithms, each of the independent variables in the training dataset were examined to see if they appeared to have a relationship with the dependent variable. Although it is persuasive to use all the independent variables in the modelling process, it is often more beneficial to use those variables which are the best in predicting the dependent variable. Reducing the number of variables reduces the likelihood of multi-collinearity. Multi-collinearity occurs when two or more independent variables are highly correlated with each other. This results in the model building algorithm not knowing which independent variable to include in the analysis. Therefore, we look for independent variables that are highly correlated with the dependent variable and which are not highly correlated with any other independent variable. A set of experiments to identify multi-collinearity problem in significant variables is explained in chapter 7.

5.7 Fuzzy variables

Definition: A fuzzy variable is characterised by triplet $(X, U, R(X; u))$, in which X is the name of the variable, U is a universe of discourse (finite or infinite set); u is a generic name for the elements of U and $R(X; u)$ is a fuzzy subset of U which represents a fuzzy restriction on the values if u imposed by X. The universe of discourse defines a set of upper and lower bounds for the values of the fuzzy sets used to describe the concepts of the fuzzy variable

A fuzzy variable is a variable with (labels of) fuzzy sets as its values. Complete definition of fuzzy variable can be seen in footnote of next page. Reason for introduction of fuzziness in the variables is due to incomplete knowledge, missing values in observed data, and rejection of some observed data. Additionally, guessing of non-observed relations among variables leads to fuzziness in the environment. Similarly, risk of re-admission can be classified into high, not very

high, somewhat low or low risk of re-admission. The transition from high risk of re-admission to low risk of re-admission can be shown by gradual transition from high to low, which can be shown by fuzzy set with degree of membership. These features and the ability to deal with linguistic terms could explain the reason of applying fuzzy methods in healthcare problems.

Risk of re-admission can be shown with the help of membership function $\mu_A(x)$, where $\mu_A(x)$ denotes the membership function MF of degree of membership of X in fuzzy set A, where X is readmission of a patient and $\mu_A(x)$ gives membership function for X which represents "risk of readmission". Membership functions have been defined in detail in chapter 3. Risk of re-admission can be categorized into "high", "medium" or "low" risk re-admission, which can be represented by membership functions. . In our research, we have used triangular and trapezoidal membership function. Triangular and trapezoidal membership functions for low, medium or high risk of re-admission is shown as in Figures 13 to Figure 15.

$$\mu_L(x) = \begin{cases} 0 & for\ x = 0 \\ 1 & for\ x = 1 \\ 1 - x & for\ 0 < x < 1 \end{cases} \qquad 1.7 \Bigg\} \text{ for low risk of readmission}$$

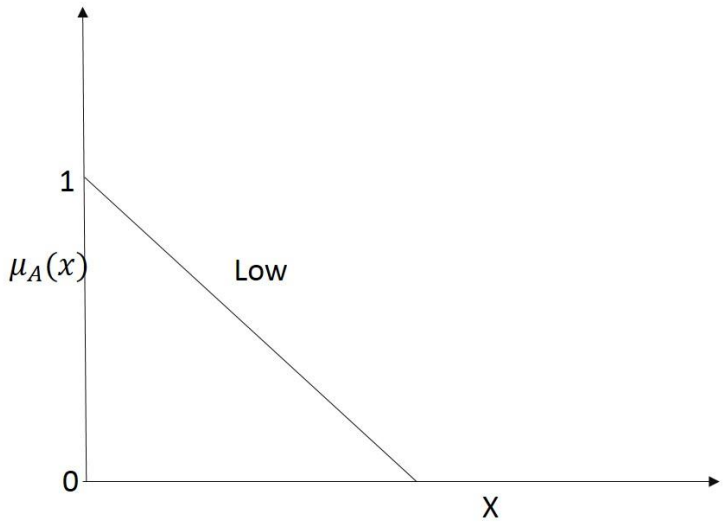

Figure 13 Triangular membership function for low risk of re-admission

1

$$\mu_M(x) = \begin{Bmatrix} 0 & for\ x \leq x_1 \\ \dfrac{x - x_1}{x_2 - x_1} & for\ x_1 \leq x \leq x_2 \\ \dfrac{x_3 - x}{x_3 - x_2} & for\ x_2 < x < x_3 \\ 0 & x \geq x_3 \end{Bmatrix} \quad (1.8)$$

For medium risk of readmission

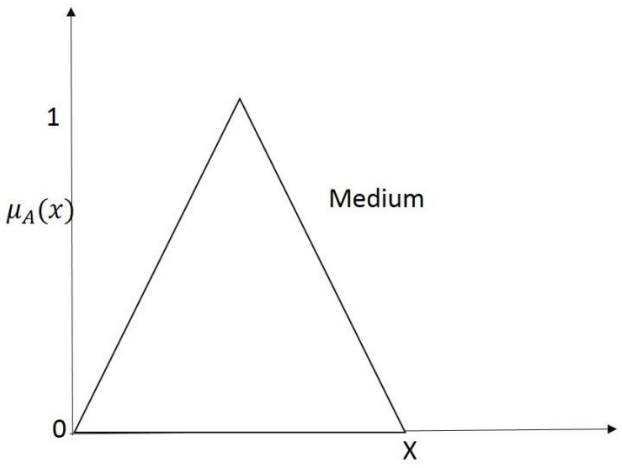

Figure 14 Triangular membership function for medium risk of re-admission.

$$\mu_H(x) = \begin{Bmatrix} 1 & for\ x = 0 \\ 0 & for\ x = 1 \\ 1 - x & for\ 0 < x < 1 \end{Bmatrix}$$ For high risk of readmission

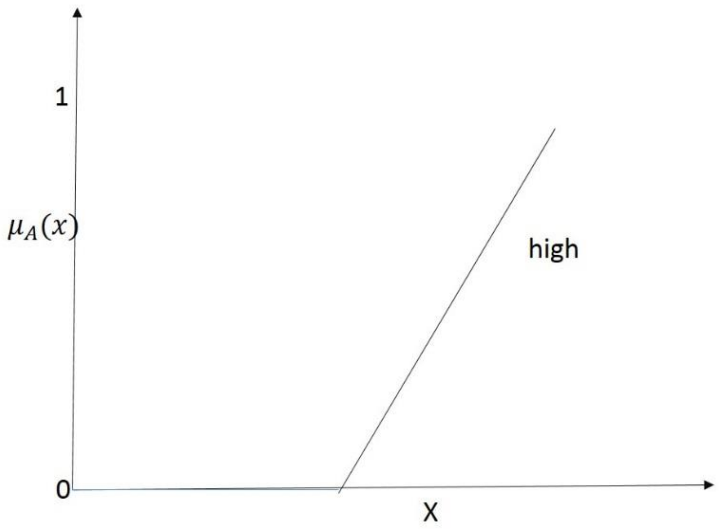

Figure 15 Triangular membership function for high risk of re-admission

Dependent variable (Risk of readmission) is fuzzified using triangular membership function. Readmitted patients can be at high, medium or low risk. Membership function defined for patients at risk of readmission is e.q(1.7), e.q(1.8), and e.q(1.9).

For any value of X, where X is readmission and $\mu_A(x)$ denotes degree of membership function MF of X in fuzzy set A. $\mu_A(x)$ gives "risk of readmission". Risk is categorized into "high, medium or low "risk of readmission is shown above with membership functions as $\mu_H(x)$, $\mu_M(x)$ and $\mu_L(x)$. Membership function for high risk of readmission means $\mu_H(x)$ value is 0 (Not readmitted) has a membership value of 1 while outcome value that equals 1(Readmitted) has a membership value of 0. An outcome X in between[0,1] has a membership value that equals to 1-X. Similary, membership function can be defined for medium and low risk of readmission.

$$\mu_L(x; x_1, x_2, x_3) = max\left(\min\left(\frac{x - x_1}{x_2 - x_1}, 1, \frac{x_3 - x}{x_3 - x_2}\right), 0\right) \quad (1.9)$$

$$\mu_M(x; x_2, x_3, x_4) = max\left(\min\left(\frac{x - x_2}{x_3 - x_2}, \quad \frac{x_4 - x}{x_4 - x_3}\right), 0\right) \quad (1.10)$$

$$\mu_H(x; x_3, x_4, x_5) = max\left(\min\left(\frac{x - x_3}{x_4 - x_3}, 1, \frac{x_5 - x}{x_5 - x_4}\right), 0\right) \quad (1.11)$$

For given values of $(x; x_1, x_2, x_3, x_4, x_5)$ as (0.4; 0.0,0.2,0.5,0.7,0.8), and using e.q(1.9), e.q(1.10) and e.q(1.11) we get:

$$\mu_L(x; x_1, x_2, x_3) = max\left(\min\left(\frac{0.4 - 0.0}{0.2 - 0.1}, 1, \frac{0.5 - 0.4}{0.5 - 0.2}\right), 0\right) = 0.33$$

$$\mu_M(x; x_2, x_3, x_4) = max\left(\min\left(\frac{0.4 - 0.2}{0.5 - 0.2}, \frac{0.7 - 0.4}{0.7 - 0.5}\right), 0\right) = 0.66$$

$$\mu_M(x; x_2, x_3, x_4) = max\left(\min\left(\frac{0.4 - 0.5}{0.7 - 0.5}, 1, \frac{0.8 - 0.4}{0.8 - 0.7}\right), 0\right) = 0$$

Our problem is to fuzzificate all real values of the variable x. In case of risk of readmission for a given value of X, for example X_n, risk of readmission can belong

to one or more MF. We calculate the value of Y for each of the membership function to which X_n belong. The value of MF lies between 0 and 1. For example, in our case we have membership functions for high, medium and low risk of readmission for a given value of X_n. The degrees of membership to each MF (Y values) for X_n as 0.4 can be, for 0.33 for the MF for low risk of readmission and 0.4 for MF for medium risk of readmission. Similarly, we can fuzzificate all values for any variable

5.8 Summary

A model that is poorly specified due to missing important input variables or inclusion of unnecessary variables will have a low fitting. Therefore, data analysis is an important part of building a model. Data preparation and manipulation plays an important role to increase the predictive power of a model.

In our proposed framework, a model is established by using more than one independent variable and one response variable. A large number of variables could lead to problems and poor model performance. Therefore, performing analysis on data has provided us a way to select significant independent variables for our model. The idea is to include as many significant independent variables as possible and at the same time checking for correlation among input variables. The next chapter describes the framework designed and developed for predicting likelihood of re-admitting patients to the hospital. Independent variables assessed during the data analysis act as input variables for predicting response variable (risk of re-admission) in our framework.

Chapter 6

6. Development of a Framework adapting Fuzzy Regression Method

6.1 Introduction

This chapter describes a framework to predict patients at likelihood of re-admission in the next 12 months of their discharge. The proposed framework is designed and developed to capture the uncertain nature of risk of re-admission. Figure 16 gives the proposed approach to represent the uncertainty in risk of readmission. Uncertainty in risk of readmission may be due to two reasons. 1. Uncertainty in output or response variable (risk of readmission). 2. Uncertain relationship between health system input variables and output variable. Fuzzy regression method with triangular or trapezoidal membership function is used to show uncertainty in decision making. Our proposed framework is based on theoretical study of fuzzy regression methods, fuzzy sets and interval-valued fuzzy numbers. Therefore, before developing the framework, preliminary theory of fuzzy regression methods is described. Our model is developed based on our proposed framework to determine significant input variables in order to predict patients at likelihood of readmission within 12 month of discharge. Our model is implemented in an algorithm to identify patients and stratify them into various risk threshold levels. A fuzzy regression method is adapted to develop our algorithm that uses the selected input variables to evaluate the response variable (risk of readmission). We have also tested our model to identify patients at risk of readmission within 30 days, 6 months and 12 months. During the development of the model, each of the independent variables in the framework was examined to select potential predictor variables. Potential predictor variables will be used as covariates in the development of an algorithm to predict high risk individuals. Input variables to the model may be uncertain due to lack of information or missing values of the dataset. Therefore, it becomes important to handle uncertain input health system variables.

Problem may occur when two or more independent variables are highly correlated with each other. This makes it difficult to identify which independent variables to include in the analysis. In order to handle multi collinearity problem concept of Interval value fuzzy numbers (IVFN) is used. In the algorithm1.1 we have shown how to handle multi collinearity problem.

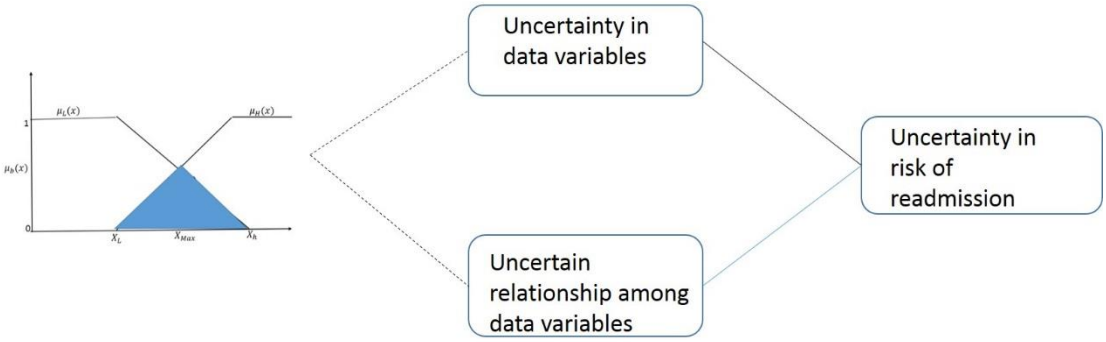

Figure 16 The proposed approach for capturing uncertainty in risk of admission

Overall chapter 6 is divided into various sections. First section describes the preliminary theory for the development of a framework. In second section, a framework is described to select significant independent variables from health system variables. Our novel algorithm that uses predictor variables to identify patients at risk of re-admission is described in next section (section 6.3.2). A part of algorithm to handle uncertain data and multi-collinearity problem is shown in more detail in the algorithm (1.1). In section 6.3.5, model validation techniques are discussed which are discussed in more detail in chapter 8. Finally, summary of the chapter is given.

6.2 Preliminary theory of proposed framework

Fuzzy sets introduced by Zadeh in 1965 were used to represent and manipulate data and information which possess uncertainties. Fuzzy numbers are numbers that can be defined in linguistic terms, for example ‘around 50 percent’, ‘a relatively high’, ‘very tall’. (Shapiro, 2005). A fuzzy number can be a triangular fuzzy number, or trapezoidal fuzzy number which is explained in detail in chapter 3. The general characteristic of the fuzzy number can be represented as a membership function as shown in figure 17. A membership function is a curve that defines how each

point in the input space is mapped to a membership value (or degree of membership) between 0 and 1. In our thesis, we have considered the triangular and the trapezoidal membership function. Examples of membership function diagrams are shown in detail in chapter 4.

6.2.1 Fuzzy Linear Regression

In recent years, there is growing literature that formalizes the linear regression model in a fuzzy domain, in which model parameters and/or data are uncertain. The linear regression model is the most frequently used form in regression analysis for expressing the relationship between one or more explanatory variables and response. In the classical statistical technique, the observations (response variable or the explanatory variables) are required to follow certain probability distributions (Billings, et al., 2006; Austin, 2007; Kim, et al., 1996). Regression analysis is a fundamental method to model crisp relationship between the dependent and independent variables based on given data. On the other hand, for fuzzy output case (Tanaka , 1987)proposed a possibilistic regression approach. Fuzzy regression analysis is a possibilistic regression analysis which is based on possibility concepts. Possibility regression analysis uses a fuzzy linear system as a regression model. The advantage of Tanaka's possibilistic regression is in its simplicity in computation. Fuzzy linear regression analysis was proposed by Tanaka et al to determine the fuzzy linear relationship:

$$\bar{Y} = A_0 + A_1X_1 + \ldots . A_nX_n \qquad (1.12)$$

Where $\bar{Y}$ is the fuzzy output,A_j, $j = 0,1,2,,3,,,,n$ is a fuzzy coefficient, and $X = X_i,,,,,X_n$is an n-dimensional non-fuzzy input vector. The fuzzy components were assumed to be a triangular fuzzy numbers (TFNs). Coefficients of the equation can be shown by a membership function (MF),$\mu_A(a)$, a representation of which is shown in figure 17

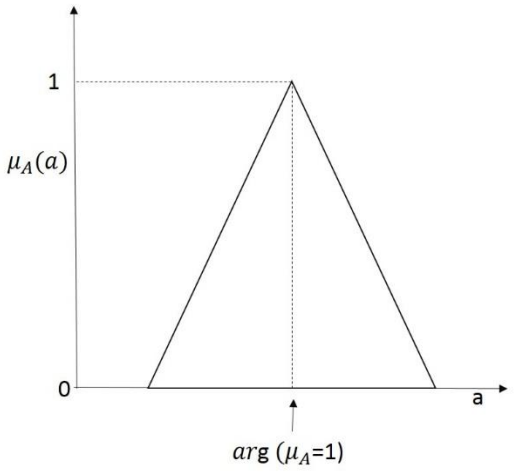

Figure 17 Triangular Membership Function

The basic idea of possibilistic regression was to minimize the total spread of the membership function, subject to including all the given data (Shapiro, 2005).

There may be two general ways to develop a fuzzy regression model. We focus in our study on the models where the input variables are crisp and response variable is fuzzy.

Case 1: Independent variables (x) are numbers (=crisp) and response variable(y) is fuzzy.

Solving fuzzy regression method using linear programming

The fuzzy linear regression model can now be re-written as:

$$\bar{Y} = (\alpha_0, c_0) + (\alpha_0, c_0)X_0 + (\alpha_1, c_1)X_1 + \ldots . (\alpha_n, c_n)X_n$$

In the following linear programming problem (LP), estimate

$$A_j = (\alpha_j, c_j)$$

$$Minimize\ J = \sum_{j=0}^{n}(c_j \sum_{i=1}^{n} x_{ij})$$

$$Subject\ to \sum_{j=0}^{n} \alpha_j x_{ij} + (1-h) \sum_{j=0}^{n} (c_{jx}\, x_{ij} \geq\ \ Y_i$$

$$\sum_{j=0}^{n} \alpha_j x_{ij} + (1-h) \sum_{j=0}^{n} (c_{jx}\, x_{ij} \leq\ \ Y_i$$

Where $a_j\ \in R, c_j \geq 0\ , j = 1,2, \ldots . n$

$$x_{i0} = 1, i = 1,2, \ldots . k$$

$$0 < h < 1$$

Where J the total fuzzinessof is the fuzzy regression model and h value is the threshold level that determines the degree of fitness of the fuzzy linear model to its data.The above Linear Programming problem was solved for different input variable sets.The central value αj together with the corresponding half-width cj of each fuzzy variable obtained for inputvariable (Age) is evaluated using MATLAB

6.2.2 Fuzzy Logistic Regression

In contrast to fuzzy linear regression, there have been few articles on fuzzy non-linear regression. It is mentioned that over the last few years there have been a few attempts to combine fuzzy regression models and logistic regression. (Pourahmada, 2011) introduced and applied a new term called possibilistic odds and then, developed a possibilistic-based regression in which the observations of the dependent variables are reported as a real number in [0, 1] representing the possibility of belonging to category 1. (Takemura, 2004) used a fuzzy logistic

regression model in which input data, output data and parameters were all represented by Linear Regression fuzzy numbers.

In some practical studies, the response variable is measured by linguistic terms such as very low, low, medium, high, and very high, rather than by precise numbers. In traditional statistics, in order to regress a binary response variable with two categories on a set of explanatory variables$X = x_1, x_2, , , , x_n$, a binary logistic regression model was used. When the response variable is evaluated by linguistic terms, the binary response variable cannot be defined precisely. Therefore, probability of success cannot be calculated. A novel approach to this problem, which was initially proposed by (Taheri et al., 2008) and (Pourahmada, 2011) is to rate the possibility of success for each observation by defining a proper fuzzy number for each term of the linguistic variable. These fuzzy numbers should be defined in such a way that their support covers the whole range of [0, 1]. In our research, we have mostly focused on fuzzy linear regression methods, but in future work fuzzy logistic regression method can be adapted.

6.3 Framework for identifying patients at risk of re-admission

The proposed framework has been designed and developed to meet the specific objectives defined in chapter 1. Specific objectives of the research are to develop a framework thatidentifies patients at risk of readmission within 12 months. Our framework includes a model to identify high risk individuals and stratifying patients into high, medium and low risk of re-admission. To achieve our second objective as stated in chapter 1, we have designed and developed a novel algorithm which adapts fuzzy regression method to identify patients at high, medium and low risk of re-admission. As shown in figure 2 above from PARR, patients are stratified into high, medium and low risk of re-admission with crisp boundaries. However, the boundaries of risk of re-admission are not crisp, and targeting individuals at borderlines can avoid hospital readmissions. It is often difficult to represent degree of certainty of patients at high, medium or low risk of re-admission. With proper healthcare interventions, patients can move from high to medium and medium to low risk boundaries. Risk of re-admission with non-crisp boundaries can be represented with fuzzy membership function

As risk of re-admission is uncertain, methods adapting fuzzy regression method provide a good approach for dealing with such type of uncertainty. Risk factors (clinical, social, patients' characteristics and demographic characteristics) are used in determining patients at high risk of re-admission. A novel algorithm is used to identify risk factors, and risk of re-admission of a patient is evaluated as a response variable. The various risk factors for patients' re-admission as derived from literature review are shown in table 4.

Risk factors for re-admission of a patient	**Independent or Response variable**
Age	Independent variable
Severity of illness	Independent variable
Type of care	Independent variable
Morbidity/comorbidity	Independent variable
Functional disability	Independent variable
Prior admission	Independent variable/Response variable

Table 4 Risk factors for risk of re-admission.

Additionally, the relationship between predictor variables (patient characteristic and disease) and response variable (risk of re-admission) is uncertain. Linguistic nature of risk of re-admission can also be defined by the fuzzy set for risk of re-admission. Both uncertainty in risk of readmission, and uncertain relationship between response & predictor variables can be modelled by fuzzy regression methods. If input variables are also fuzzified, the algorithm may be computationally intensive in evaluating uncertain relationship among variables. As a result, a large of number of fuzzy rules will make the algorithm computationally slow. Therefore, in our algorithm we have fuzzified only the response variable. In our approach, we have modelled crisp inputs with non-crisp output. Description of health system variables and selection of independent & dependent variables is given in next section. In addition, there may be a problem of multi-collinearity among various

variables. An algorithm adapting fuzzy regression method with Interval valued fuzzy numbers (IVFN) is proposed to treat multi-collinearity problem. The framework for identifying patients at risk of re-admission is depicted in figure 18.

Figure 18 depicts the framework using health system variables to identify patients at risk of readmission to the hospital. A model is developed within our framework to identify potential predictors for readmission. Using fuzzy regression method we develop our model to identify and stratify readmitted patients at various risk threshold levels. Identification of patients at risk of re-admission is an expected outcome of the algorithm. Before design and development of algorithm, significant independent variables are checked to see if they have relationship with risk of re-admission. Although, we have a large set of input variables, it is beneficial for us to determine significant independent variables. Significant predictor variables are used as potential covariates for our algorithm. As an output of the algorithm, significant independent variables as risk factors responsible for risk of re-admission are also determined.

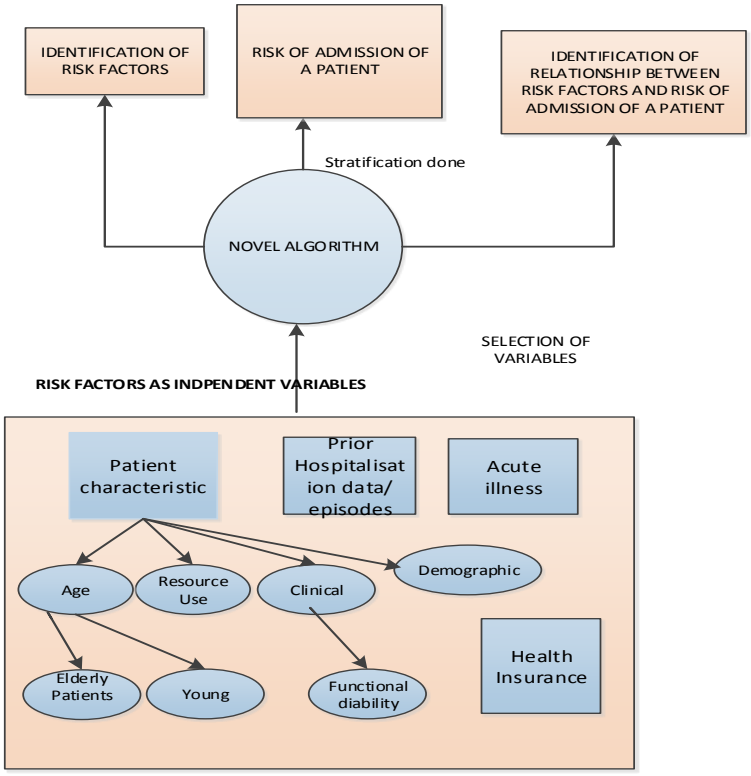

Figure 18 The framework for identifying patients at risk of re-admission

The framework shown above is implemented and verified with these three important steps:

1. Data preparation & pre-processing.

2. Design and development of the model implemented in the proposed algorithm.

3. Assessing the model performance

Each of the steps are discussed in more detail in next section. Model performance is described briefly in this chapter and is discussed in more detail in chapter 7.

6.3.1 Data Preparation and processing

This section contains a brief outline of how the data were prepared and processed for this thesis. The data source used in this project was Hospital Episode Statistics (HES) data containing individual episodes of care over financial years (1999/2000 to 2004/2005). Each record in the inpatient part of the database is a finished consultant episode. Approximately 3.5 million inpatient admissions that started and finished in 2004/2005 within England were included in our study. The extracted

data for the 109,243 records included the variables which gave us the information like the age at start of admission, gender, triggering admission date, discharge date and diagnostic conditions, Charlson comorbidity index(CCSI). The following binary variable ("Readmission_12") was created for each of the sample of 109,243 emergency admissions that started and ended in 2004/2005 to see if the patients had a subsequent emergency admission within 12 months of the discharge date of the triggering admission. A set of variables based on patient's prior utilisation were created and these data were combined with data on patient's characteristics and diagnostic conditions. Predictive models are generally 'trained' on a data set consisting of dependent variables (Readmitted patients in hospitals) and a range of independent variables from record of patient in previous years. For performing statistical analysis, a chi-square univariate analysis was carried out to determine which patient characteristics and health outcomes had significant impact for hospital re-admission. We also look for independent variables that appear to have relationship with dependent variable. Table 5 below shows the final independent variables included in the analysis. These variables were used as inputs to the fuzzy regression model to predict re-admission within 12 months.

Patient characteristic variables for e.g age, gender and ethnicity are normally correct. Admission date We have used diferent variables for our model for e.g age, alcohol abuse, anaemia, angina, drug abuse, average number of episodes per emergency admission spell etc as shown in table 5. When a patient is diagnosed with an alcohol or drug related disorder, the diagnosis is often complex, as these conditions are susceptible to both psychological and physiological signs, symptoms and manifestations and cormobidities.

ICD-10-CM codes provided for these diagnoses is based on ICD-9-CM codes, which may be complex. In ICD-9 the details focused more on timelineof the patient's use of the alcohol or drug involved, while ICD-10 –CM requires understanding of pyschological or behavioural impact. This may lead to huge variability in NHS trust. Various diagnosis variables are based on ICD-10 codes.Disease presence and diagnostic history are based on ICD codes in any diagnostic field (primary or secondary) in discharge data. Therefore, they are subject to variability. ICD-10

provides with combination codes with dependence codes, thereby inducing inconsistency in the dataset. Reference conditions are given by HRG codes, which specifies the presence of complicated medical conditions

Independent variable	Independent variable
Age 75 and over at admission	Drug abuse
Alcohol abuse	Injury from fall
Anaemia	Ischaemic heart disease
Angina	Mild Liver disease
Atrial fibrillation	Number of emergency admissions within the previous 3 years
Average number of episodes per emergency admission spell	Number of emergency admissions within the previous 6 months
Average number of episodes per non emergency admission spell	Number of nonemergency admissions within the previous 3 years
Cancer	Reference condition in the previous 3 years
Congestive heart failure (CHF)	Renal Failure
Chronic obstructive pulmonary disease (COPD)	Respiratory infection
Connective tissue disease/rheumatoid arthritis	Severity index total score
Development disabilities	Sickle cell disease
Diabetes	White

Table 5 List of variables included in the model.

6.3.1.1 Fuzzification of response variable

As stated above dependent variable “Re-admission” is created from HES variables to see if the patients had subsequent readmisision in the next 12 months. In the classical set theory we could say that re-admission of patient is either admitted (1) or not (0), but for risk of re-admission we consider degree of membership to a fuzzy set. A ‘membership function’ --a possibility distribution-- is the main aspect of fuzzy set theory. The function describes the possibility (degree) to which a measurable value “re-admission” belongs to a particular linguistic term like “risk of “re-admission”. Fuzzy set that represents “risk of re-admission” does not comprise only two elements “yes” (1) or “no”(0) as it does in the case of classical set theory. Rather, it constitutes an array or range of points, a possibility distribution that fills the interval between “0” and “1”. A range of values for “risk of re-admission” can be [high, medium, low] that lies in the interval of [0,1] Schematically, the main step is fuzzification which is establishing a membership function with links between some values that are not necessarily ordinal, but are associated with a phenomenon. Similarly, fuzzification of re-admission is done by establishing a triangular membership or trapezoidal membership function which maps ordinal values of risk of re-admission as shown in figure 19 and figure 20. Due to their simple formulas and computational efficiency, both triangular and trapezoidal MFs are used extensively. Due to simplicity and computationally efficient, we have used trapezoidal membership function uncluding traingular membership function for our research.

Membership function $\mu_L(x)$ is for low risk of re-admission, $\mu_M(x)$ is for medium risk of re-admission, and $\mu_H(x)$ is for high risk of re-admission, where x is readmission and $\mu(x)$ is a membership function for x, representing risk of readmission. X_h is the value of X, which corresponds to membership function for high risk of readmission, X_L is the value of X, which corresponds to low risk of readmission, and X_{Max} is the mode of X, which corresponds to the maximum value of membership functions. Similarly, membership functions for high, medium or low risk of re-admission are used for trapezoidal membership function.

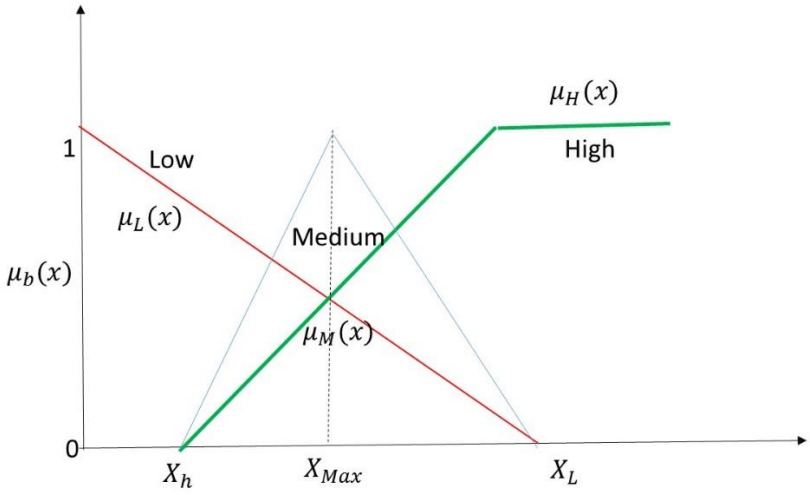

Figure 19 Triangular membership function for [high, medium and low] risk of re-admission.

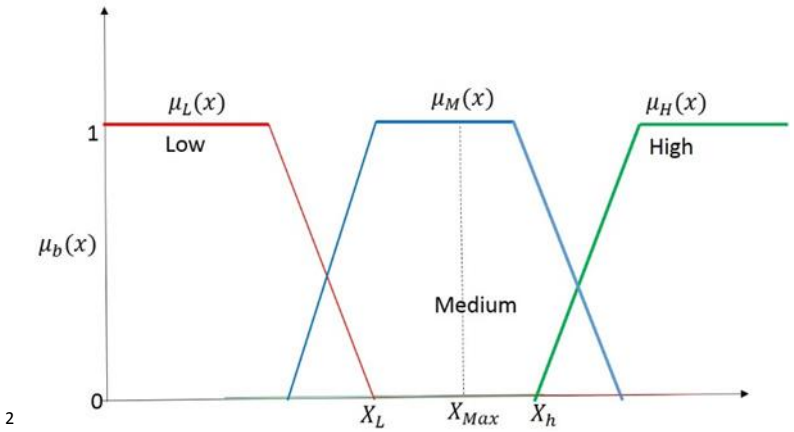

Figure 20 Trapezoidal membership function for [high, medium and low] risk of re-admission

[2] The process of translating the measured numerical values into fuzzy linguistic values is called fuzzification.

6.3.2 Algorithm Outline and Descriptions (I)

Hospital risk of re-admission of patients tend to be uncertain due to unexpected variations in the data, patients' characteristics, demographic variables and health system variables. Significant independent variables (patients' characteristics and demographic characteristics) from health care data helps in identification of risk factors responsible for predicting patients at risk of readmission. Traditional models are available to identify patients at likelihood of readmission within 12 month of discharge. However, these models fail to deal with uncertainty in the risk of readmission of a patient. In addition, uncertain relationship exists between input variable and response variable. In our algorithm, prediction of "risk of re-admission" is predicted by evaluating the relationship between "risk of re-admission" and "risk factors". Uncertainty in model prediction of response variable may arise from a number of sources including estimation of input values and interpretation of predicted outcome of a model. Of these, uncertainties due to estimation of input values can be handled with regression method. Uncertainties that arise from the interpretation of the predicted outcome can be dealt with fuzzy sets. However, handling uncertain input health variables with fuzzy membership functions may increase the computational complexity of the algorithm. Therefore, we have dealt with crisp inputs variables. In our approach, we have adapted fuzzy regression method to handle crisp inputs and fuzzy output. Details of the algorithm is shown in figure 21, and are described here.

The first stage is data preparation and processing stage. It is divided into different steps including identification of membership function and handling of uncertain data. In the first stage, a fuzzy membership function is defined as a triangular or trapezoidal membership function, with risk of re-admission as a fuzzy set {high, medium or low}.

Then we deal with health system variables such as patient characteristics. Data preparation and processing is carried out for cleaning and manipulating of health system variables. As, we are dealing with risk of readmission as a response variable, we fuzzify the response variable to represent it in linguistic terms as high, medium or low risk.

In the second stage, the solution approach is more elaborate and important. At this stage, the patient at risk of re-admission is identified as response variable and other health system variables (such as patient and disease characteristics) are identified as input variables. Regression analysis is carried out to identify the relationship between and independent variables and risk of readmission. Significance of independent variables is determining by p-value. If p-value is less than 0.05, then we can reject the null hypothesis that no relationship exists between independent variables and dependent variables. Using the Chi square analysis, relationship between independent and response variables are identified. The relationship cannot be mapped always linearly.

In traditional regression framework, several procedures have been suggested for choosing set of explanatory variables. A model could be established by using more than one independent variable. However, a large number of independent variables could lead to a problem of correlation among variables which makes the prediction model multi-collinear. The idea is to include as many independent variables as possible but at the same time avoiding co-linearity. Traditional models fail to deal with uncertainty in these risk factors. In addition, uncertain relationship exists between data variables. In many cases, non-linear relationship exists between predictor and outcome variables. In our research, we have only dealt with fuzzy linear regression analysis. When reducing a non-linear relationship to linear we need to follow following steps

1. Identify the functional relationship in a form containing three terms.
2. Make one of the terms to be a constant.
3. Remove unknown constants from the coeffiecient of one of the variable terms.
4. Compare with the standard from of linear equation Y= mX+C, where m is the slope of the line and c is the vertical intercept value.

In order to handle the problems of data uncertainty and of multi-collinearity, an approach with an Interval value fuzzy number is shown in fig 22. This can be explained in three different stages which is explained in algorithm 1.1

In the third stage, Fuzzy linear regression analysis is carried out by solving fuzzy regression coefficient using linear programming in R. The upper and lower bound of the response variable is evaluated with the help of fuzzy regression method. This forms a fuzzy set ($\bar{y}$,$\underline{y}$) and value lies in the fuzzy set {high risk, medium risk, low risk} of re-admission.

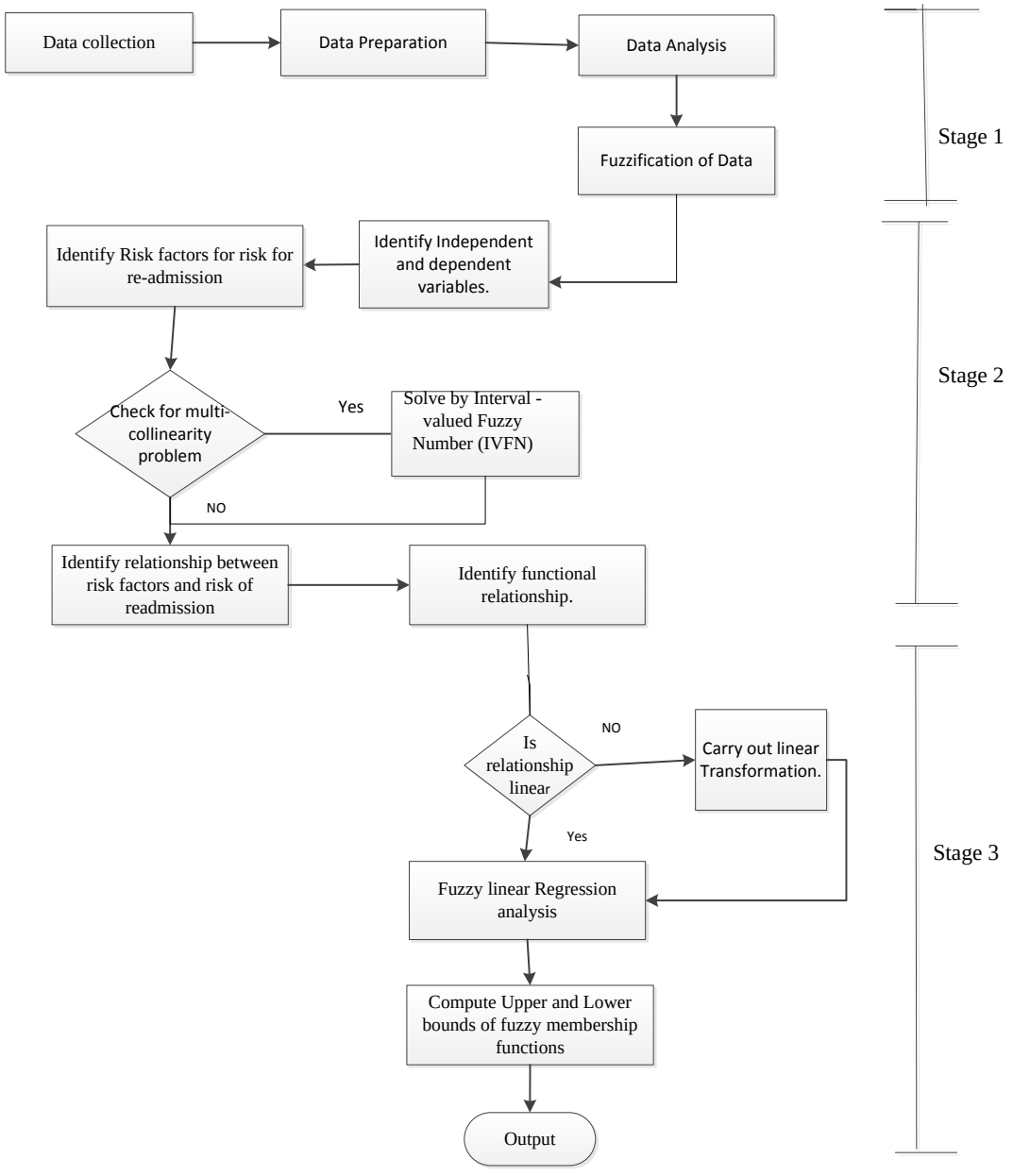

Figure 21 An algorithm adapting fuzzy regression method

6.3.3 Checking Multi-collinearity

In a traditional regression framework, several procedures have been suggested for choosing a set of explanatory variables. A model could be established by using more than one independent variable. A large number of independent variables could lead to a problem of correlation among variables which makes the prediction model multi-collinear. The problem of multi-collinearity may make the model more complex. In order to make computation simple, we have avoided co-linearity problem in health variables by using interval-valued fuzzy numbers.

In order to handle data uncertainty and multi-collinearity problem, an approach with an interval valued fuzzy number is shown in fig 22. This can be explained in three different stages in algorithm 1.1

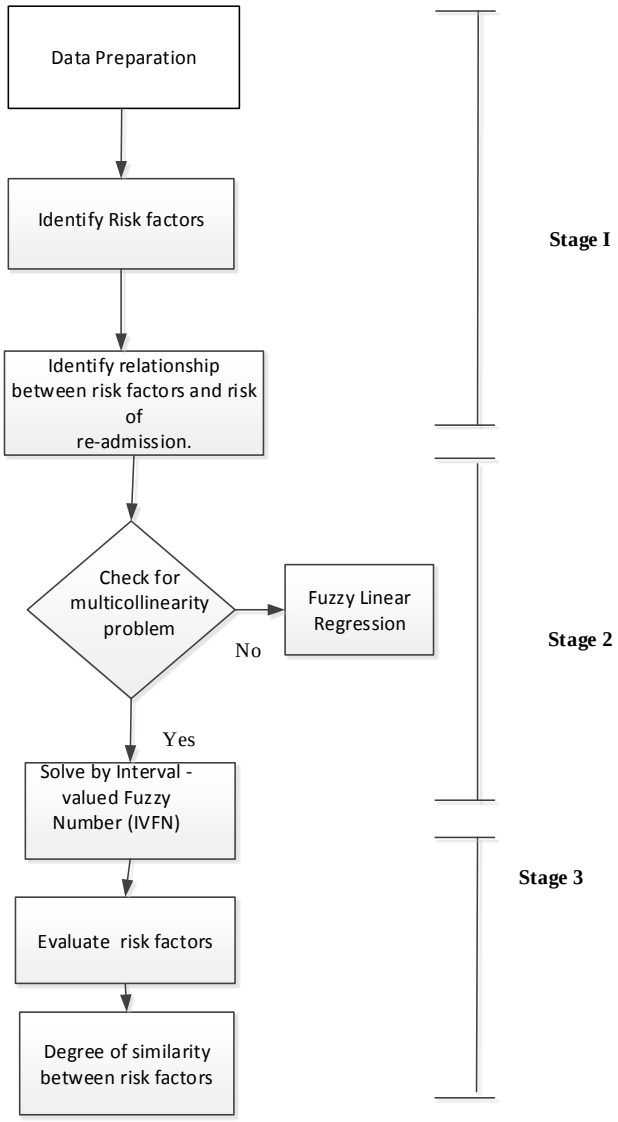

Figure 22 A Proposed approach to handle multi-collinearity problem between risk factors

6.3.4 Handle multi collinearity problem

Stage I: In this stage, preparation of data is done for fuzzy regression analysis. Data preparation includes identification of uncertainty in data variables (such as patient's characteristics and disease characteristics). Significant independent and response variables are identified for fuzzy regression approach. Fuzzy membership function is identified and fuzzification of data is done to make data suitable for carrying out fuzzy regression analysis. We would like to include all the variables in

our model, but a large of number of variables may increase the problem of multi-collinearity. Reducing the number of variables may help in increasing the predictive power of model. Risk factors responsible for risk of re-admission are identified by analysing significant input variables.

Stage II: This stage involves steps to check for multi-collinearity problem. If multi-collinearity does not occur in data variables then the fuzzy linear regression analysis is carried out. The problem arises when there is multi-collinearity among risk factors. This can be solved by using Interval value fuzzy numbers with triangular or trapezoidal membership functions. Risk factors are identified using fuzzy regression model with interval-value fuzzy numbers.

Stage III: Finally, area of two interval valued fuzzy numbers has been used to calculate the degree of similarity between lower fuzzy numbers and upper fuzzy numbers in the Interval - value fuzzy numbers. Once the degree of similarity is found, it can be used in ranking of risk factors at an individual level.

6.3.5 Solving Fuzzy regression methods

According to Tanaka's possibilistic regression the response variable Y can be written as (Arulchinnappan & Rajendran, 2011; Rosma, et al., 2008; Fialho, et al., 2012)

$$\bar{Y} = A_0 + A_1X_1 + \ldots . A_nX_n \qquad (1.12)$$

Where $\bar{Y}$ is the fuzzy output,and $X = X_i \ldots \ldots X_n$ is the real valued input vector of independent variables and each regression coefficient $A_j, j = 0,1, \ldots . . n$ was assumed to be an symmetric triangular fuzzy number with centre α_j and half width $c_j, c_j > 0$. (Arulchinnappan & Rajendran, 2011; Rosma, et al., 2007; Fialho, et al., 2012)

The Fuzzy Regression equation is considered for a single input variable. This can be further extended to multiple risk factors with multiple variables.

$$Y = A_0 + A_1X_1 \ldots \ldots \ldots . . (1.13)$$

Where $Y = \left(\underline{y}, \overline{y}\right), A_0 = \left(\underline{a}, \overline{a}\right), A_1 = \left(\underline{b}, \overline{b}\right)$ and $\underline{y}$ is the lower bound and $\overline{y}$ is the upper bound of fuzzy regression equation $\left(\underline{a}, \overline{a}\right)$ and$\left(\underline{b}, \overline{b}\right)$ are regression

coefficients of lower and upper bound regression equations. Regressing the equation and then multiplying the whole equation by $\sum_{i=1}^{n} X_i$ in order to solve the equation and find values for upper bound and lower bound regression coefficients.

$$\sum_{i=1}^{n} Y_i = n\,\overline{A}_0 + \overline{A}_1 \sum_{i=1}^{n} X_i \qquad (1.14)$$

$$\sum_{i-1}^{n} X_i\, Y_i = \overline{A}_0 \sum_{i=1}^{n} X_i + \overline{A}_1 \sum_{i=1}^{n} X_i \qquad (1.15)$$

Substituting the values of the upper and lower bound values in equation (1.14) gives equation (1.15) and (1.16).

$$\sum_{i=1}^{n} Y_i = n\,\overline{a} + \overline{b} \sum_{i=1}^{n} X_i \qquad (1.15)$$

$$\sum_{i=1}^{n} Y_i = n\,\underline{a} + \underline{b} \sum_{i=1}^{n} X_i \qquad (1.16)$$

Substituting the values of the upper and lower bound values in equation (1.15) gives equation (1.17) and (1.18).

$$\sum_{i-1}^{n} X_i\, Y_i = \overline{a} \sum_{i=1}^{n} X_i + \overline{b} \sum_{i=1}^{n} X_i \qquad (1.17)$$

$$\sum_{i-1}^{n} X_i\, Y_i = \underline{a} \sum_{i=1}^{n} X_i + \underline{b} \sum_{i=1}^{n} X_i \qquad (1.18)$$

Solving these equations and calculating the values of Lower bound and upper bound values using a program gives the values as:

Substituting the values of the upper and lower bounds derived from above program in equation (1.17) and (1.18), will give upper and lower bound equations that identifies the relationship between risk factors and risk of re-admission of a patient.

Similarly, this equation can be extended for multiple risk factors and regression coefficients as$A_0, A_1, A_2, \ldots\ldots A_n$ can be evaluated for upper and lower bounds. The value of the fuzzy output lies in the set$\left(\underline{y}, \overline{y}\right)$.

Solving Interval Valued Fuzzy Numbers (IVFNs)

In the proposed methodology, the theory of interval valued fuzzy sets deals with uncertainty and multi-collinearity in risk factors. IVFN have been adopted to handle uncertainties in risk factors arising from incomplete and imprecise information (Hung & Yang, 2006). In this approach, interval valued fuzzy sets are used to represent degree of membership of a function, a similarity measure to calculate degree of similarity between Interval value fuzzy numbers is used. In order to find degree of similarity, fuzzy weighted mean method and interval valued fuzzy number are used in this study. On the basis of similarity measure calculated between IVFN, it helps in analysis and ranking of risk factors in uncertain environment

An interval value fuzzy set

$$C = \{(x, [\mu_C^L(x), \mu_C^U(x)]) | \, x \epsilon X\} \qquad (1.19)$$

Where $0 \leq \mu_C^L(x) \leq \mu_C^U(x) \leq 1$ and the membership grade $\mu_C(x)$ of the element x belongs to the interval valued fuzzy set C which can be represented by the interval.

where $\mu_C^U(x)$ denotes upper bound of IVFN and $\mu_C^L(x)$ denotes lower bound of IVFN. Assuming two interval valued fuzzy sets as a set A and B has two elements, where the other one as upper fuzzy number A^U and the lower fuzzy number as A^L. Similarly, for fuzzy set B, the lower fuzzy number is B^L and the other upper fuzzy number is B^U. IVFN fuzzy sets can be represented as

$$A = [A^L, A^U] = [(a_1^L, a_2^L, a_3^L, a_4^L; w_A^L), (a_1^U, a_2^U, a_3^U, a_4^U; w_A^U)] \qquad (1.20)$$

$$B = [B^L, B^U] = [(b_1^L, b_2^L, b_3^L, b_4^L; w_B^L), (b_1^U, b_2^U, b_3^U, b_4^U; w_B^U)] \qquad (1.21)$$

Where $0 \leq w_A^L \leq w_A^U \leq 1$

$0 \leq w_B^L \leq w_B^U \leq 1$

Where $(a_1^L \leq a_2^L \leq a_3^L \leq a_4^L), a_1^U \leq a_2^U \leq a_3^U \leq a_4^U$,

If $a_1^L = a_1^U$

$a_2^L = a_2^U$

$a_3^L = a_3^U$

$a_4^L = a_4^U$

And

$b_1^L = b_1^U$

$b_2^L = b_2^U$

$b_3^L = b_3^U$

$b_4^L = b_4^U$

And

$w_A^L = w_A^U = w_A$

$w_B^L = w_B^U = w_B$

then the interval valued fuzzy numbers can be regarded as generalized fuzzy numbers. The multi-collinearity problem can be solved based on the degree of similarity between two interval valued fuzzy numbers. Degree of similarity $S(A, B)$ can be evaluated by calculating the areas $A(A)$and $A(B)$ of the trapezoidal fuzzy numbers. Area of the two fuzzy numbers can be evaluated as where a_1, a_2, a_3, a_4, are element of fuzzy number A and b_1, b, b_3, b_4, elements of fuzzy number B

$$A(A) = \frac{1}{2} w_A (a_3 - a_2 + a_4 - a_1) \qquad (1.22)$$

$$A(B) = \frac{1}{2} w_B (b_3 - b_2 + b_4 - b_1) \qquad (1.23)$$

The larger the value of $S(A, B)$, the more the similarity measure between two trapezoidal fuzzy numbers.

6.3.6 Outlier Treatment

Outliers in general may represent problematic data, but in our case the outliers are values. Domain knowledge is very important in determining how one should handle these extreme values. Unusual or extreme observations (outliers) for interval or continuous (non-binary) variables are usually removed from training datasets prior

to the application of algorithms to ensure that the models are built using stable and consistent data. Outliers are not removed from the validation dataset as each of the models should be tested on actual data to determine their true performance.

6.3.7 Assessing the model performance

A framework which utilises model that adapts fuzzy regression method to capture uncertainty in risk of re-admission is proposed in this chapter. As explained above, fuzzy regression method was developed with significant variables extracted from HES data. Our proposed framework deals with uncertain nature of "risk of re-admission", and uncertain relationship between risk of readmission and input variables. Traditional methods such as logistic regression have been developed to predict patients at high risk of re-admission. In classical regression model, patient is either readmitted or not, which is depicted as yes (1) or no (0). It is beneficial to validate the generalization of the proposed algorithm with traditional methods, and compare results of the proposed approach and classical models.

6.4 Summary

The proposed framework in this chapter helps to account for the uncertain nature of risk of re-admission. Because of nature and ill-defined boundaries of risk bands, this approach does allow the user to identify individuals at high risk of re-admission. Fuzzy regression method is chosen due to its flexibility in handling uncertain data. Patients at risk of re-admission could be identified with consideration of significant variables. Risk scores can be evaluated and thresholds can be set at higher levels for patients who have a history of previous admissions and are at risk of future re-admissions which is shown in chapter 8. Descriptions of the prediction measurements used such as accuracy, discriminations, calibration and area under the receiver operating characteristic curve are described in sections of chapter 8.

The fuzzy regression model was experimented on Hospital episode statistics dataset as a part of the validation exercise to check on the generalization of the algorithm in predicting patients at risk of re-admission which is shown in chapter 8.

Chapter 7

7. Experiments for model adapting fuzzy regression method.

7.1 Introduction

The main aim of study is to identify people at risk of re-admission on whom use of direct intensive resources may improve health outcomes. The issue is that a small number of patients could be classified as high risk and use a large amount of resources. The identification of readmitted patients may also mask any variation in the severity of condition and the quality of care provided for them. Risk of re-admission of a patient is uncertain because it can take the values other than 0 and 1. As stated above, risk of re-admissions can be stratified into high, medium or low risk of readmission with ill-defined boundaries. A problem may arise in prediction due to the uncertain nature of risk of re-admission of a patient. Risk of readmission is an output variable that depends on the significant input variables.

We have conducted various experiments before starting with actual implementation of fuzzy regression method. Throughout this thesis, we are using different terms as uncertainty in risk of readmission, uncertainty in health variables, and uncertain relationship among dependent& independent variables. We have experimented to test for uncertainty in healthcare data variables, and uncertainty in decision making to make sure that fuzzy regression method can be implemented.In this chapter, we will demonstrate the various experiments carried out to show the feasibility of our algorithm.

Before implementing fuzzy regression method, we have carried out experiments to make sure that our proposed methodology is valid. Initially, we started with MATLAB but later on, we faced difficulty in finding licensed package for fuzzy regression method in MATLAB. A part of the algorithm could be implemented with MATLAB, where we can represent input variables and output variables with membership function.

This chapter lists all the experiments carried out in order to validate our proposed method. The set of experiments conducted are as follows:

1. To represent the uncertainty in risk of re-admission.
2. To handle uncertain health system variables.
3. To assess risk factors for patients at risk of re-admission.
4. To develop and test algorithm adapting fuzzy regression method with significant input variables.
5. To conduct an experiment to compare different models.

In the first experiment (section 7.2) surface viewer plot is generated as shown in figure 23 which is helpful in understanding how the system is going to behave for the entire range of values in the input space. This experiment implements part of the algorithm, and shows how the system behaves for entire range of values. It is not clear, or is uncertain, to say what kind of relationship exists between dependent and independent variables. For e.g., increase in input variable may lead to increase in output variable, whereas in other cases with more than one input variable surface viewer may show another kind of relationship. This kind of relationship becomes complex to understand. In first experiment, we have explained in detail for uncertainty in risk of re-admission, and uncertain relationship in health system variables.

We have also experimented on uncertainty in health variables. It becomes confusing to understand which variable is fuzzified - input or output. Input variables can be represented with triangular or trapezoidal membership functions. Although, it may be interesting to fuzzify input variables, we found no reason to do so. Also, if we fuzzify one input variable then why not the other one. If all input variables are fuzzified, then it increases the computational complexity of the algorithm. It becomes ambiguous to understand which rules to include in the analysis. Therefore, in our model we have only fuzzified response variable.

In the second experiment (section 7.3) P-Plot is used to understand whether data variables are normally distributed or not. The plot is a graph of the empirical

Cumulative Distributed Function (CDF) values plotted against the theoretical CDF values. It may happen that input variables are not-normally distributed, which indicates that correlation and regression cannot always be applied to the dataset.

In order to validate, in experiment 4 (Section 7.5), we have implemented our model in an algorithm using fuzzy regression method. In order to carry out full implementation, we have to go for licensed package. Therefore, we have decided to start implementation in R due to easily available open source packages for fuzzy regression method and its ease of use. We have explained in detail in last experiment the final implementation of fuzzy regression method in R.

In experiment 5(Section 7.6) we have done experiment to compare fuzzy regression method with other traditional methods (logistic regression, neural network, decision tree). Comparison of the prediction performance of different data mining methods is evaluated with the help of ROC curves. Implementation of these methods is done with the help of open source R packages.More detailed validation of the different methods is given in chapter 8.

In our research, we have used a machine learning technique based on the concept of fuzzy regression method to develop our framework. The use of the above method is appropriate since we are dealing with uncertain data, and since the explanatory variables interact in uncertain manners. To evaluate the performance of our algorithm we tested the model adapting fuzzy regression method on HES dataset.

Experiments are conducted using HES data obtained from NHS information centre for health and social care for the period 1999/2000 to 2005/2006. Records were extracted of all NHS hospital admissions in England for emergency inpatient admissions that started and ended between 1/04/2004 and 31/03/2005. The next emergency admission which was within 12 months of the discharge date of the triggering admission for these patients was also extracted. All of the variables were derived for use in predictive modelling methods. Chi square and Regression analysis was carried to see the significance of each and every independent variable. The variables tested were based on a broad range of measures used in the algorithm

which predicts readmission in the following year. These measures included the number of admissions to the hospital; number of episodes per spell in prior admissions; prior utilisation of hospital resources in the last 12 months; and a range of diagnostic categories. The reduced numbers of variables ultimately included in this model were selected based on their impact on overall model performance. Potential covariates for the risk prediction model that uses HES are age, emergency admission, history of admissions in the prior five years, the number of admissions in the past 6 & 12 months prior to current admission, and severity index score.

Although it is appealing to use all independent variables in the development of model, problem arises when there is multi-collinearity within data variables.Risk prediction model could encompass a set of risk factors such as age, severity of illness, comorbidity, prior admission and other factors for e.g., type of care and functional disability. Fuzzy regression method with interval-valued fuzzy numbers is used to solve this problem of multi-collinearity. Degree of similarity between risk factors is found using interval-valued fuzzy numbers. In experiment 3 (section 6.3), we have assessed various risk factors using triangular and trapezoidal fuzzy numbers.

Predicted outcome “re-admission” is a binary outcome and can be evaluated using logistic regression, but “risk of re-admission”, due to its linguistic behaviour, can be modelled by fuzzy regression method.

7.2 Experiment to represent uncertainty in risk of re-admission

The following experiment is conducted to understand uncertainty in risk of readmission of a patient. Conceptually, we have described above that “risk of readmission” is uncertain, as boundaries of risk stratification are not crisp. Response variable (risk of re-admission) will be represented by a range of values from high, medium to low risk. From our dataset, we have derived variables as Readmission_12. In our model, we are interested in “risk of readmission”, which is a fuzzy variable. Fuzzy membership function is appropriate to represent uncertain nature of predicted outcome. Therefore, representing risk of readmission with triangular membership function is useful. Additionally, there can be uncertain

relationship between input health variables and “risk of re-admission”. In this experiment, we have tested membership function of response variable for varying input variables.

7.2.2 Methodology

The framework to predict patients at high risk of re-admission is described in chapter 6. The proposed framework includes an algorithm to identify likelihood of re-admission. In this experiment, we have tried to test part of the algorithm with MATLAB. This section of the algorithm steps is described in figure 21.It shows the algorithmic steps for the experiment carried out. Details of the algorithm are described in different stages as given below.

The first stage of the experiment is data preparation and processing. Significant Iindependent variables are selected .The dependent variable derived from data set is Readmission_12. A sample of trained dataset is used for the analysis. Risk of readmission can be defined in linguistic terms [high, medium or low], and can be represented with a fuzzy set [high, medium or low]. We are using triangular membership function for this experiment. Response variable is fuzzified using a fuzzy membership function. Input variables used are age, number of previous admissions, and severity of illness. **In second stage** of the experiment, the risk of re-admission is the response variable. Selected health system variables (such as patient and disease characteristics) are treated as input variables. Relationship is evaluated between response variable and independent variables. The type of relationship among variables is also checked for linearity. **In the third stage,** fuzzy linear regression analysis is carried out by solving fuzzy regression coefficient using the linear programming method.

We did not have package of fuzzy regression method in MATLAB therefore, evaluation of method is done in JAVA and then program is imported into MATLAB. The upper and lower bound of the response variable is evaluated with the help of fuzzy regression method which is shown above in equation (1.17) and equation (1.18) in chapter 6. This forms a fuzzy set $\left(\underline{y}, \overline{y}\right)$and its value lies in the fuzzy set [0,1] of risk of re-admission.

7.2.3 Results

Risk of re-admission can be represented by fuzzy membership function. Uncertainty in risk of re-admission is shown with the help of triangular membership function. It shows membership function for fuzzy set [high, medium and low] risk of re-admission. Output variable (risk of re-admission) which lies in range of [0, 1] varies depending on input variables as shown in surface viewer plot.

Figure 23 shows the surface viewer plot of the fuzzy membership function for the input variables affecting the output variable. The surface viewer plot is evaluated for one to two input variables as age, severity of illness and output variable as risk of re-admission. The number of inputs can be increased from one to four. The x and y axes plot the input variables such as age and severity of illness and response variable is on z axis as risk of re-admission. The plot described how the varying input variables affect response variable of risk of re-admission. Output variable (risk of re-admission), shown in figure 23, can have range of values in a set [0, 1] for varying input variables

Figure 24 represents the fuzzy membership function depending on the input variables. It represents the triangular fuzzy membership function for training the data sets. It shows that the variation in input variables affects the output of the function. It also shows the rules for the various input factors. The decision will depend on the variation in input factors. The red line corresponding to the input factors shown can be moved to compute the effect on output. As, we have conducted this experiment for a sample of a small dataset, we have tested for the uncertain relationship between response variable and input variables. Input variables are represented by membership function, but we did not find it useful. There is no reason to represent some of the input variables with membership function. Additionally, varying the input variables results in large number of fuzzy rules making the algorithm computationally slow.

While evaluating fuzzy regression method we have fuzzified the response variable. In this experiment, output is represented by bold blue lines. The fuzzy set with range of values can be defined as {high, medium or low} risk of re-admission. The

output that we receive as a result of fuzzification can be mapped to values in a fuzzy set as shown with bold lines in figure 24. It shows the output variable for two and three input variables.. It gives the degree of membership for the output variable (risk of readmission). Once the input variables are fed into the system, we have to define the membership function for the output variable. For a given value of x, the value of response variable can be evaluated. Degrees of membership function for output variable will lie between 0 and 1. For e.g if we consider three values of membership function , the degree of membership to each MF(Y values) for input variables can be for example 0.6 for the MF low, 0.4 for MF normal and likewise. In case, we have more than one input, the degree of membership for output variable will be minimum value of degree of membership for different inputs. The blue lines represent the risk of readmissionThis is only representing the membership function.The intersection point between different triangles is calculated.The values of the bold lines of each MF is evaluated for the membership function, which gives us the value for risk of readmission. The output values are evaluated by calculating the point at which a line would balance the triangles.

We used MATLAB for another experiment (based on IVFNs and explained later). However, results obtained were not very clear. Moreover, to get the clear identification of results, fuzzy regression method package was not available. Hence we did not use MATLAB for carrying out further analysis.

Considering the above, we have implemented fuzzy regression method in R which is explained in detail in experiment 6.5.

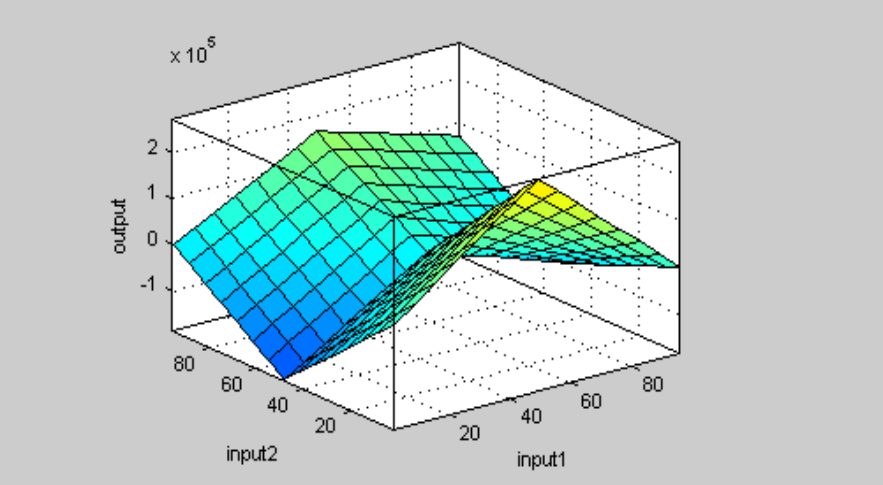

Figure 23 shows the surface viewer plot of the input-output surface of the fuzzy.

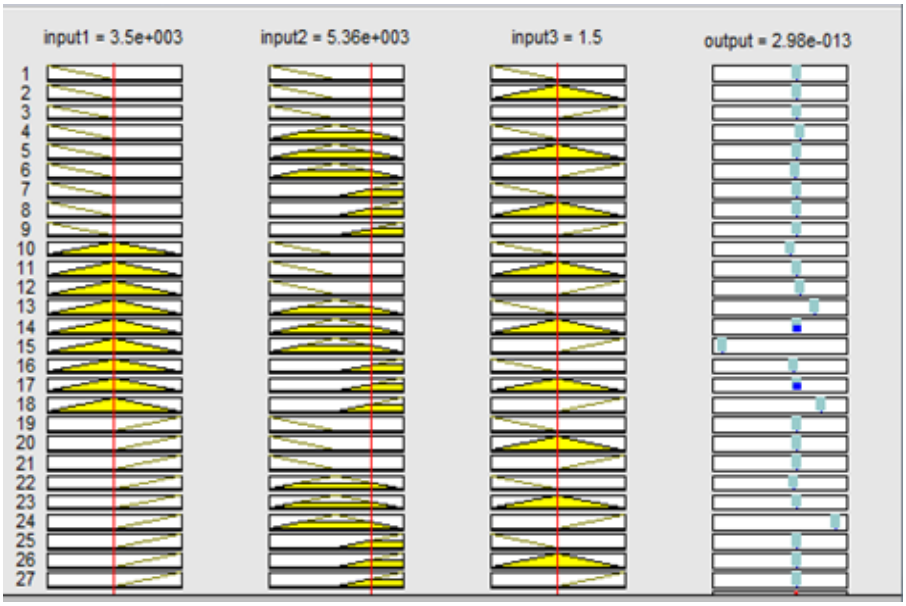

Figure 24 The triangular membership function plots after training dataset

7.3 Experiment to understand nature of health system variables.

A sample of data from the HES dataset was extracted. Before applying any algorithm, it was important to understand the nature of the data variables. This experiment was conducted to check whether data is normally distributed or not. If the data variables are not normally distributed then traditional techniques such as 1-sample t test, 2-sample t-test, and one –way ANOVA cannot be always applied. Additionally, it gives us the idea for the type of relationship that exists among health system variables. Most of the relationship among health system variables

can be captured by linear relationship. Not all data can be captured by linear relationships. The type of relationships among variables can be linear or non-linear. It may be complex to understand the type of relationship between independent variable and dependent variable.

Regression Analysis can be used to assess the associations between risk of re-admission and independent variables. The test for normal distribution is conducted through Probability-Probability Plot. Purpose of p-p plots is to test for whether data is normally distributed, and to check for outliers in the dataset. (P-P) plot will be approximately linear if the specified theoretical distribution is the correct model.

7.3.2 Method and Results

The test for normal distribution is conducted through Probability-Probability Plot (P-P) on a sample of the extracted dataset. The plot is a graph of the empirical CDF values plotted against the theoretical CDF values. The plot will be approximately linear if the specified theoretical distribution is the correct model. The P-P Plot indicated that all data is not normally distributed. Insufficient data discrimination – and therefore an insufficient number of data values might become the reason for uncertainty in the data variables. It is apparent by regression analysis that strong and statistically significant relationship exists between some variables, but other variables have a weak relationship. Also, correlation between some variables is not statistically significant. Figure 25 shows the P-P Plot of severity of illness variable, which is deviated from the straight line. It shows that linear relationship does not exist and data cannot be captured by just correlation and regression. Similarly, it is shown for P-P Plot for age variable in figure 26. In figure 27, P- P Plot of comorbidity variable is shown, which is U-shaped. This type of relationship also cannot be captured by just correlation and regression. As the data is not normally distributed, correlation and regression techniques may not be always appropriate. This experiment helped us in clarifying that on this dataset, liner regression techniques cannot be always applied. Alternately, other type of data mining methods (such as Logistic regression, neural network, decision tree, fuzzy regression methods) can be considered for such type of data. Considering the above, we have adapted fuzzy

regression method to estimate the relationship between health system variables. A summary of results is provided in table 6 with significance value. Independent variables are significant if p-value <0.05. As, we could see that all variables have p-value <0.05, therefore they are considred to be significant for predicting risk of readmission.

Parameter	**Standard Error**	**Significance**
Age 75 plus at admission **(0)**	0.0121	<.0001
Average number of episodes per emergency admission spell	0.0153	<.0001
Number of emergency admissions within the previous 5 years	0.00633	<.0001
Number of emergency admissions within the previous 6 months	0.0225	<.0001
Reference condition in the previous 5 years **(0)**	0.0135	<.0001
Severity Index	0.00891	<.0001
White **(0)**	0.0102	<.0001
Alcohol **(0)**	0.0254	<.0001
Cancer **(0)**	0.0198	<.0001
CTDRA **(0)**	0.0341	<.0001
Development disability **(0)**	0.0568	0.0019
Diabetes **(0)**	0.0179	0.0068
Drug abuse **(0)**	0.0452	0.0002
Injury from fall **(0)**	0.0166	<.0001
Mild Liver Disease **(0)**	0.0619	0.0163
Number of emergency admissions within the previous 6 months	0.0276	<.0001
Number of non-emergency admissions within the previous 12 months	0.0100	<.0001
Congenital disability **(0)**	0.0327	<.0001

Table 6 Significant independent variables for predicting re-admission within 12 months.

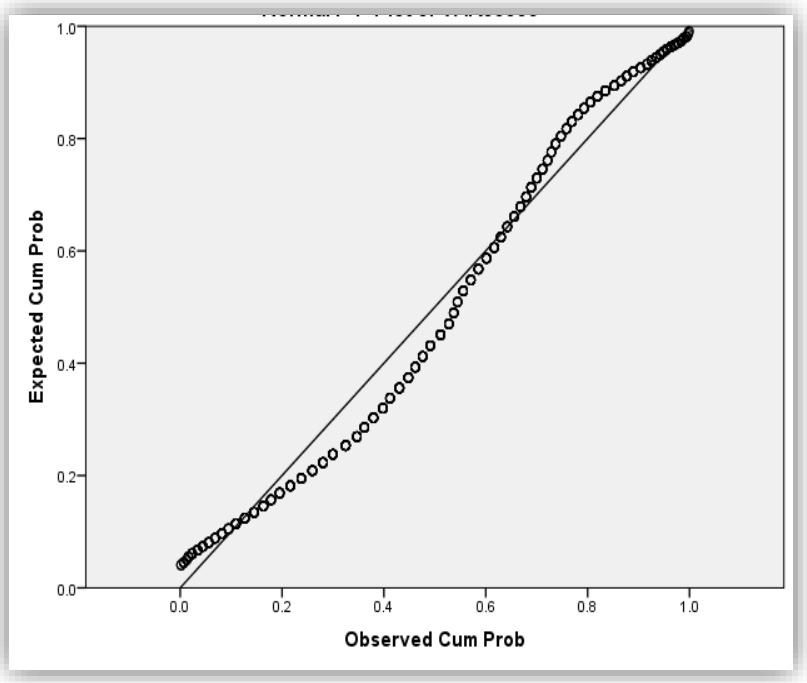

Figure 25The P-Plot of severity of illness variable

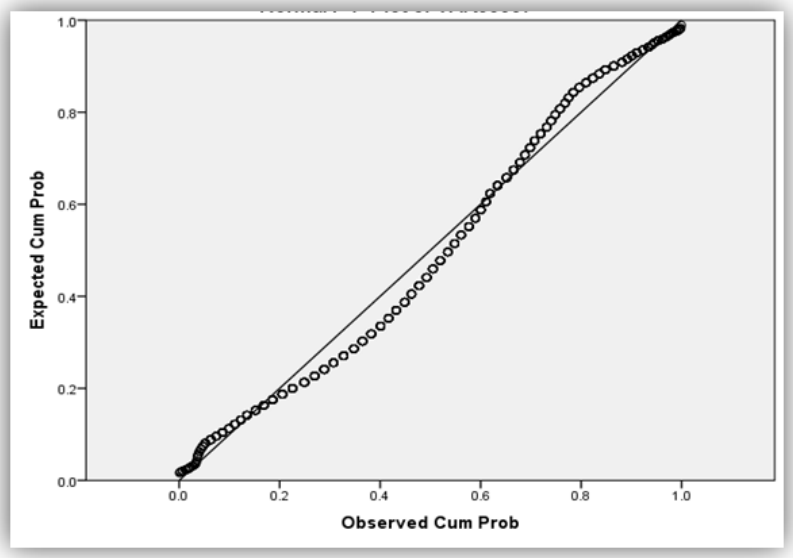

Figure 26 The P-Plot of age variable

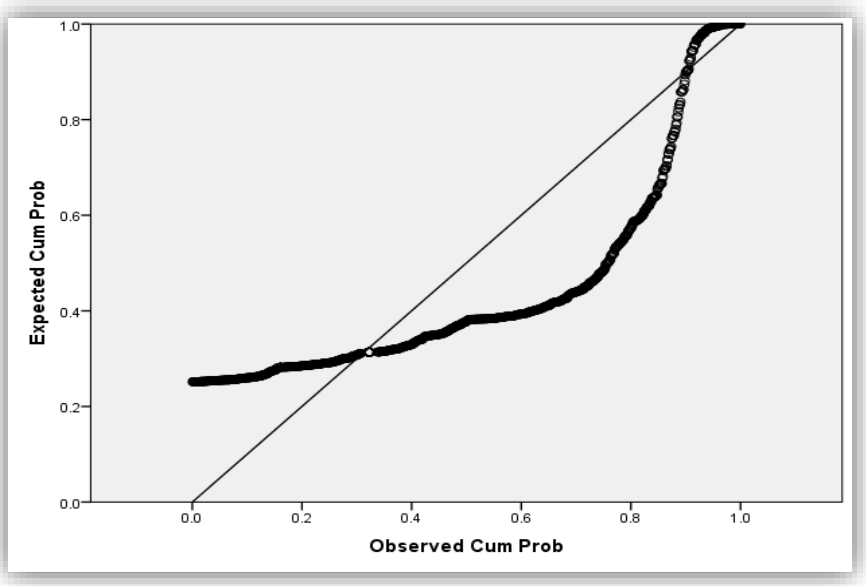

Figure 27 The P-Plot of comorbidity variable.

7.4 Experiment for assessing risk factors using Interval-Valued Fuzzy Numbers (IVFNs)

In risk prediction model, independent variables are used as input variables for predicting patients at risk of readmission. A large number of input variables may improve the predictive power of the model, but may increase the problem of multi-collinearity in data variables. Risk factors include age, severity of illness, reference condition, etc. in risk prediction model. Fuzzy regression method with interval-valued fuzzy numbers is an attempt to treat uncertainty and multi-collinearity in risk factors. Risk factors can be assessed and ranked by finding degree of similarity in interval-valued fuzzy numbers. We have used triangular and trapezoidal membership functions for assessing risk factors. This experiment is an attempt to assess risk factors using interval-valued fuzzy numbers.

7.4.2 Methodology and Results

In order to handle uncertainties in an effective manner (Zadeh, 1965), developed the theory of fuzzy sets and utilized this theory to model uncertainty or lack of

knowledge in decision making for a variety of problems. The fuzzy set is an effective method to deal with imprecise linguistic terms based on a range of values. Decision makers find it difficult to handle uncertainties arising due to lack of knowledge or incomplete information. Experts often find it difficult to identify an opinion as a number in the interval [0, 1]. Therefore, to represent degree of certainty of opinions, fuzzy sets with interval valued fuzzy numbers (IVFN) is used in the analysis. Risk of readmission is represented by fuzzy membership functions. Membership functions are triangular or trapezoidal membership functions with symmetric shape and equal spread. In order to handle multi-collinearity problem and assess similarity between risk factors an interval value fuzzy number is used. Interval value fuzzy numbers is shown with trapezoidal membership functions and are used to solve multi-collinearity problem between independent variables. Degree of similarity between fuzzy numbers is calculated by evaluating the area of fuzzy numbers.

A sample of trained dataset from HES dataset is used for this experiment. Risk factors are selected as input variables for implementation in this experiment. We have used fuzzy toolbox in MATLAB to conduct this experiment. In this experiment, we have shown trapezoidal fuzzy membership function to treat multi collinearity problem. Plots show the trapezoidal membership function. In this experiment, we have shown the methodology and how we can represent risk of re-admission as trapezoidal membership function.

The proposed method provides a useful way to handle risk factors in a complex and uncertain environment. Fig 28 and Fig 29 represent the trapezoidal fuzzy membership function. These figures show how the variation in input variables affects the output of the function. This is one of the experiment in which we are assessing degree of similarity between input variables using interval-value fuzzy number. Results show how input variables can be shown with membership functions. With trapezoidal membership function as value of μ approaches 1 slope increases and as value of μ approaches 0 the value of slope decreases. The decreasing slope of membership function help us in understanding of patients at low risk of readmission and increasing slope at high risk of readmission.

The fuzzy set encompasses a range of output values, and bold lines represent a value from the fuzzy set. The output for membership function is depicted by bold blue lines. The range of values from high to low risk of re-admission can be mapped to values with the help of bold lines.

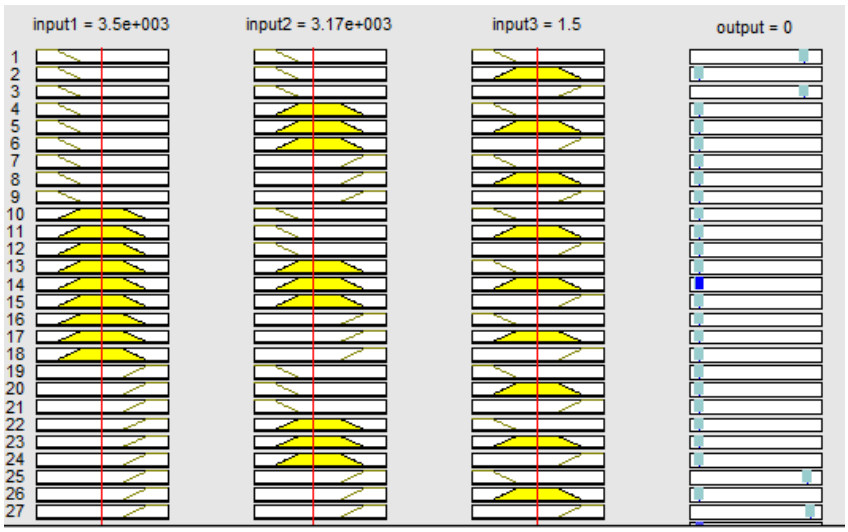

Figure 28 Trapezoidal membership function plot for the inputs specified

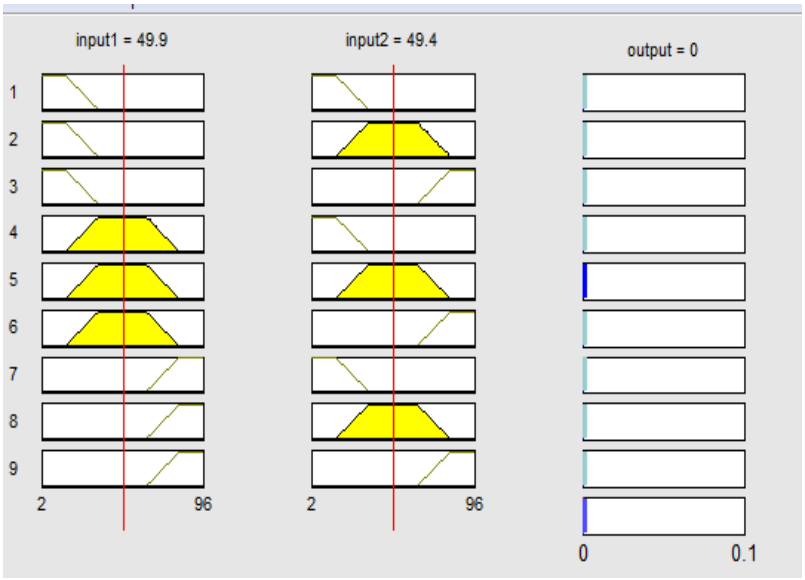

Figure 29 Trapezoidal membership function plot for the inputs specified

Potential covariates or predictors associated with patients' re-admission are assessed. Uncertainty in decision making can be represented with the help of

interval valued fuzzy numbers. The relationship between risk of re-admission and risk factors associated with it are shown with the help of regression equations and degree of similarity is evaluated with the help of areas of interval-valued fuzzy numbers (IVFN) as shown in algorithm 1.1. Risk factors responsible for risk of re-admission can be evaluated and ranked with the help of Interval-valued fuzzy numbers. Through this innovative fuzzy regression approach using interval valued fuzzy number, it is appropriate to deal with multi-collinearity among variables.

We have represented interval-valued fuzzy numbers with trapezoidal membership function. Full implementation of fuzzy regression method with interval-valued fuzzy number is explained in next experiment.

7.5 Experiment to develop and test adapted fuzzy regression algorithm.

The experiment on HES dataset was started with input variables fed into the system and the output was recorded and measured. The number and choice of input variables was done on the basis of analysis as explained in chapter 5. Fuzzy regression implementation is done using FRBS (Fuzzy rule-based systems) package in R based on the concept proposed by Zadeh. Fuzzy regression methods are important to tackle problem of uncertainty, and they are commonly used for identification and regression tasks. We focus on learning from data with learning methods of classification and regression.

The model in our implementation is $Y \rightarrow \int (X_1 \ldots., X_n)$ where the output function is a linear combination of the input variables. Here X_i and Y are the input and output variables. The model performs learning methods in order to construct FRBS for regression tasks from data. The frbs. learn () method and the predict () method are used to construct FRBS models and perform fuzzy regression respectively. Internal functions are invoked through frbs. learn (). In our method, we choose fuzzy variable to be “re-admission_12” and the shape of the membership function to be “TRAPEZOID”. This is depicted in figure 30. For sensitivity and specificity analysis, we have used ROCR package which is helpful in estimating performance measures and plotting these measures over a range of cut-offs.

The sensitivity, specificity, accuracy values and area under the receiver operating characteristic curve (AUC) for input variables are experimented. The descriptions on accuracy, sensitivity, and specificity are given in chapter 7. These terms are briefly reviewed here. Accuracy refers to the probability to correctly classify outcome. Sensitivity refers to the probability to predict positive outcome when true state is positive. And, finally specificity refers to the probability to predict negative outcome when true state is negative. The interpretation of results and ROC curve is detailed in chapter8.

7.5.1 Fuzzy Regression Experiment on input variable sets

The fuzzy regression models were fed with significant input variable sets. The following variables were used for the model:

1. Age group
2. Gender
3. History of previous admissions
4. Severity of Illness score
5. Charlson Comorbidity Index
6. Reference Conditions
7. Source of admission

7.5.2 Fuzzy membership function

Fuzzy membership function for “risk of re-admission” is shown in figure 30, which shows a range of values that lies in the interval of [0, 1]. The transition from high to low risk of re-admission is shown by fuzzy set with degree of membership function. High risk of re-admission is shown by membership function represented by “green line”, medium risk of re-admission by “red line” and low risk of re-admission by “blue line”.

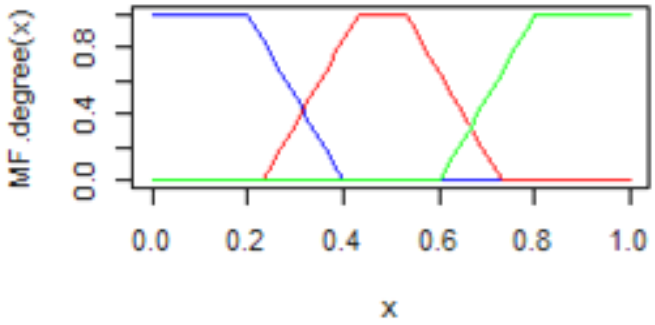

Figure 30 Trapezoidal membership function for Risk of Re-admission

The performance of a prediction model is commonly evaluated in terms of its calibration and discrimination abilities. Calibration measures how close the predictions made by a specific model are to the real outcome. This is usually done by determining whether there are any statistical significant differences between the real outcome and the predicted outcome. Discrimination, on the other hand, measures how well the two classes in the data set are separated. Calibration and discrimination were described in detail in chapter 8. These measurements are used at this point to differentiate the prediction abilities of the different models. The prediction performance of our proposed model were measured and then compared with the prediction performances of other validation models in Chapter 8.

The sensitivity, specificity, accuracy values and area under the receiver operating characteristic curve (AUC) for input variable sets experimented are presented next. The descriptions on sensitivity and specificity are given in Chapter 8. The area under the receiver operating characteristic curves was calculated. The receiver operating characteristic curve is shown in Figure 31.

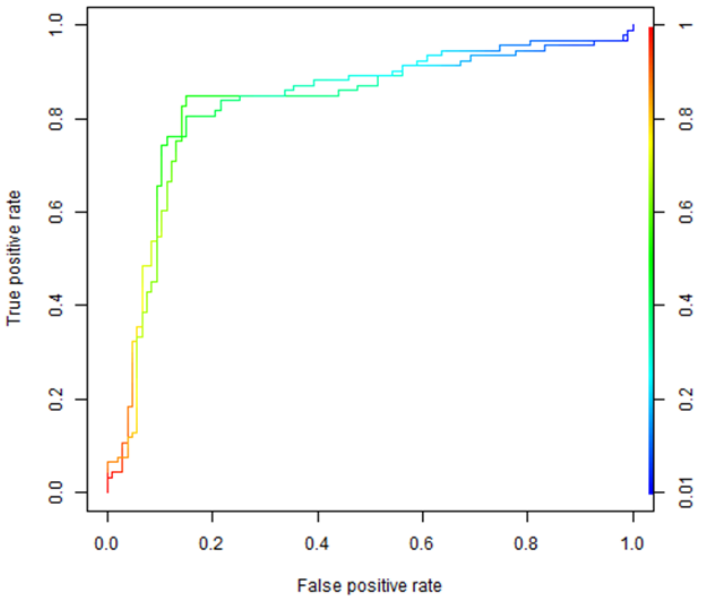

Figure 31 ROC curve for fuzzy regression method.

7.5.3 Results

The performance of the model is shown with the help of ROC curve in the figure 31 above. The receiver operating curve in the figure illustrates the trade-offs for users between sensitivity (true positives) and 1-specificity (false negatives) for the algorithm. True positives (sensitivity) and false positives (1-specificity) are evaluated at different risk scores (0-100). For our model based on fuzzy regression, sensitivity is 58.8% at the risk score of 50. The specificity of the model is the percentage of records that actually did not have re-admission within 12 months that were correctly predicted not to have re-admission by the model (true negatives). The specificity for our model is 87.4%.

In model validation, sensitivity and positive predictive value is evaluated for risk score of 40 and above as explained in more detail in chapter 8. The positive predictive value will give the percentage of records that the model predicts will have a re-admission that actually did have the re-admission.

This shows that fuzzy regression model performs well with addition of significant independent variables. The area under curve (AUC) is 0.735, which indicates a

73.5% probability that a randomly selected patient with future re-admission will receive a higher risk score than a randomly selected patient who will not have a future re-admission. More detailed analysis of sensitivity and positive predictive value is shown in chapter 7. The performance of the model is shown in terms of the percentage of patients with re-admission within 12 months. Addition of significant variables may increase the predictive power of model which will depend on the over-fitting measures of the model.

7.6 Experiment on comparison of different Models

A number of predictive models and tools have also been developed for the prediction of patients who are at high risk of re-admission (Rosma et al., 2008; Krumholz et al., 1997). These studies tend to produce conflicting results where factors associated with unplanned re-admissions vary widely in statistical significance and, as a consequence, the predictive model and the tool may not provide sufficiently accurate predictions. Most predictive models have focused on regression techniques, although there is an emerging interest in machine learning algorithm. Details of study on logistic regression, classification tree and neural network can be found in Appendix 3.

7.6.1 Methods

Prediction of outcomes is usually done using logistic regression with records of patients from HES dataset. We have implemented logistic regression method in R using glm () method. We have fitted the logistic regression model that includes both explanatory variables and response variable. For this part of analysis, we have used five years of data from 1999 to 2004 with triggering admission year data for 2004/2005. For logistic regression, potential covariates were used as input variables and the outcome predicted is modelled as a linear combination of predictor variables. Re-admission of patients which is a predictor variable, is converted to a categorical variable as “re-admission_12”. Re-admission_12 is “0” if patient is readmitted and “1” if patient is not readmitted within 12 months of the discharge date. A series of logistic regressions were conducted to identify those

variables that contributed most to the likelihood of re-admission with 12 months of discharge.

As an alternative to logistic regression methods, researchers have applied various machine learning methods, especially ANN and decision trees. Several studies have tried to compare the performance of previous predictive models (Bottle, et al., 2014) (Bottle et al., 2006) A number of studies have compared the predictive ability of decision trees with regression analysis and Artificial Neural Network.

There are a number of ways to implement machine learning methods in SAS, R and MATLAB. We have implemented decision trees in R using rpart package which includes rxDTree function. This function provides the ability to estimate decision trees on very large datasets. Decision trees provide easy to interpret models, and are helpful in predicting patients' characteristics that are associated with high risk of re-admission. Decision tree and logistic regression offer an appealing output that, unlike ANNs, shows the relation between predictor variables. Details of significant variables identified with ROC curves is depicted in chapter 8.We concentrate on the standard feed forward ANNs using neuralnet package in R. Such models are usually trained using back-propagation algorithm (Bottle et al., 2006(Bottle et al., 2014). The dataset was randomly split into training (60%) and validation dataset (40%). This was done, in the first phase to select the best model features and in a second phase to assess its algorithm. Variables were removed one after the other and performance after each removal was evaluated. The best combination of significant variables that gave the best performance were selected. Significance of independent variable for risk of readmission was evaluated based on the performance of the model. The most significant predictive variables identified were then tested on validation set. Classifier used in our model is feedforward neural network with varying number of neurons in one hidden layer and with one neuron on the output layer. In our research, we have used one hidden layer as the standard choice. The number of nodes in the hidden layer should be of the order of square root of the number of variables in the model. ANN is trained using back propagation algorithm. All independent variables were fed into the neural network, and performance of the model was evaluated. Model was trained using

backpropagation. Variables were removed one after the other and performance after each removal was evaluated. The best combination of significant variables that gave the best performance were selected. Significance of independent variable for risk of readmission was evaluated based on the performance of the model. The most significant predictive variables identified were then tested on validation set.

Neuralnet will return object of class nn which is a list containing candidate covariates responsible for predicting likelihood of re-admission.

7.6.2 Results

Traditional measures such as sensitivity and specificity are used to estimate the area under the receiver operating characteristic (AUROC) curve. The performance of the various models is compared based on ROC curve where values can range from 0.500 to 0.900. Another measure such as positive predictive value is used to predict percentage of those patients at risk of re-admission. The AUROC curve for logistic regression is 0.723. Similarly, AUROC for classification tree, and neural network is evaluated which comes out to be 0.715 and 0.699. Results of these models with their comparison is summarized in following table: Commonly used classification using AUC for a diagnostic test is summarized below.

AUC Range	**Classification**
0.9 < AUC < 1.0	Excellent
0.8 < AUC < 0.9	Good
0.6 < AUC < 0.7	Not good
0.7 < AUC < 0.8	Worthless

7.7 Summary

A classical regression technique is an estimation method which is normally used in finding crisp relationship between dependent and independent variables. The proposed approach which adapts fuzzy regression is based on Tanaka's possibilistic approach. The experiment is conducted on the HES dataset. In our methodology, a fuzzy regression model minimizes the uncertainty of the estimated values for the

dependent variables. The fuzzy regression model was experimented and proved to be useful for selecting independent variables for identifying high risk patients. The next chapter will enhance the results of these experiments. Four different models were constructed and validated on similar datasets as a part of validation exercises. Discussion and comparison of different models with significant input variables is shown in chapter 8.

Chapter 8

8. Model Validation

8.1 Introduction

The main aim of this chapter is to do model assessment and validation. This work consisted of assigning a combination of predictor variables to predict patient at risk of readmission. Models assess patients falling into one of the two classes as being readmitted or not readmitted. This could be assessed using discrimination performance. This chapter focuses on description of different performance measures such as ROC curve, risk score, risk threshold, sensitivity, specificity and accuracy of models. This is a function of true positive ratio verses false positive ratio, which is shown using AUROC curves. The true positive rate and true negative rate corresponds to the sensitivity, and specificity of the problem. In our model, true positive rate represents the case where the patient was correctly being classified as being readmitted, and true negative rate represents the patients which are being classified as being not readmitted. We have also utilised true positive and false positive rate to calculate the percentage of patients predicted to have a re-admission that were predicted correctly.

In this chapter, we have compared different models based on their performance. Several studies have tried to compare the performance of previous predictive models. A number of studies have compared the predictive ability of decision trees with regression analysis and Artificial Neural Network for multiple input variables. Predictive ability can be compared using the area under ROC. We have compared and evaluated Decision Tree, Logistic Regression, and Neural Network models with the help of ROC curves. We have also validated fuzzy regression method with the help of ROC curves. All these techniques use values of one or more independent variables to predict whether a patient had a re-admission (the dependent binary variable). For each and every technique, significant independent variables are also described. Addition of significant input variables may add to the predictive power

of the model. Detailed description of different techniques with their significant independent variables and performance of these models is given in this chapter.

8.2 Model Validation

Model validation is a major part of research work. The essential aim of this thesis is to develop a framework to predict patients at risk of re-admission. In achieving this objective, this framework utilises predictive model which adapts fuzzy regression algorithm. Fuzzy regression method is chosen due to its flexibility in handling imprecise data. As a part of the validation exercise, the fuzzy regression model was experimented on HES dataset to check the generalization of the algorithm. It is vital to prove that the proposed predictive model is an accurate and reliable model. This was done by comparing the predictive performance of the proposed model with the predictive performance of existing methods. Artificially intelligent prediction technique such as logistic regression and neural network are used for validation with fuzzy regression method. Logistic regression prediction model is used to validate the proposed model adapting fuzzy regression method since logistic regression is a commonly used method for predicting binary output. Comparison of the results on the prediction abilities of the proposed and validation models in terms of percent accuracy, discriminations and calibration are done later in this chapter. Detail descriptions of performance measures are explained in this chapter.

8.3 Model Performance

Measurements for risk prediction model used are accuracy, discriminations, calibration and area under the receiver operating characteristic curve. Discrimination is the ability of risk score to differentiate between patients who experience a re-admission event during the study and those who do not. Common measures of discrimination are sensitivity, specificity, and percent accuracy (Billings et al., 2006; Rosma et al., 2008). This measure is quantified by calculating the area under the receiver operating curve (AUROC) statistic. Advantages and drawbacks of different models including prediction ability and prediction interpretations were also analysed.

The performance of the model can be measured with a positive predictive value (PPV) for a risk threshold and an area under the ROC curve ('c-statistics'). A recent literature review of predictive risk models for 12 month re-admissions documented c-statistics ranging from 0.50 to 0.72. The area under the Receiver-Operating-Characteristic Curve (ROC) is normally used to depict the graphical representation of discrimination.

A good diagnostic test is one which has small false positive and false negative rates across a range of cut off values. The larger the area, the better the diagnostic test is. An ideal test will have an area under receiver operating characteristic (AUC) of 1 because it achieves both 100% sensitivity and 100% specificity. A bad diagnostic test is one where the only cut offs that make the false positive rate low have a high false negative rate and vice-versa.

Traditional measures of performance, such as the sensitivity, mask the potential value of models in targeting preventive interventions which are described below.

8.3.1 Risk Threshold

The sensitivity and specificity of the model can be traded off against each other by varying the threshold of risk used to define them. An overall cut-off level/threshold can be set for the full range of intervention strategies.

8.3.2 Risk Scores

Risk score is determined by observing the trade-off between true positives and positive predictive value. (Lewis, 2015) This can be explained as the false positives can be increased or decreased at the expense of increasing or decreasing false negatives. It is observed that if risk threshold decreases number of false positive increases while decreasing false negatives. On the other hand, increasing risk score threshold decreases false positives and increases false negatives. A risk score closer to 0 indicates a very low chance of re-admission, while a score closer to 100 represents a very high chance of re-admission. Any individual with a risk score of 50 or above is predicted to have a re-admission and the others (those with risk scores of less than 50) are predicted not to have a re-admission. A series of data mining algorithms were conducted to identify those variables that contributed

most to prediction of re-admissions. Prediction of re-admission can be described by creating 'risk scores' of 0-100 describing the estimated probability of re-admission within 12 months of discharge. The three techniques of logistic regression, classification trees and neural networks were used to produce a probability (between 0 and 1) of obtaining an outcome of an event. All three techniques use values of one or more independent variables to predict whether a patient had a re-admission (the dependent binary variable).

8.3.3 ROC Curve

The c-statistic, or area under the receiver operating characteristic (ROC) curve is considered in diagnostic testing (Cook, 2007). Diagnostic test characteristics of ROC curves, such as sensitivity and specificity are relevant to discriminating readmitted patients verses non-readmitted patients. Discrimination is more of interest when classification into one of the two classes of readmitted and not-readmitted is the goal. Discrimination is measured using ROC curve, or c statistics. In diagnostic setting, already determined outcome and the estimated classification are compared. The ROC curve and its associated c-static are sensitivity and specificity for each value of measure of the model. Measures of discrimination are common but they ignore random nature of outcome. In risk stratification, the outcome is not yet known, and readmitted patient's status can only be estimated as probability or risk. Calibration is a measure of how well predicted probabilities or risk match with actual observed risk. In particular novel risk factors which contribute to overall risk prediction becomes an important question. When the average predicted risk in subgroup of patient actually matches with the readmitted patient, then we say that model is well calibrated. The c-static is equivalent to the probability that measure or predicted risk is higher for readmitted patients than not-readmitted patients

We measured accuracy of predictive models in a number of ways. The plot of an ROC curve shows the sensitivity on the x-axis and '1 minus the specificity' on the y-axis. We present estimates of the area under the receiver operating characteristic (ROC) curve, which shows the trade-off between true positive (sensitivity) and false negatives (1-specificity) at all possible thresholds. The positive predictive value (PPV) is defined as the percentage of those at-risk patients identified by the model

as being at risk. Descriptions on false positive, false negative, true positive and true negative rates are summarized in next section.

The area under ROC curve (AUROC) provides a way to measure the accuracy of a diagnostic test. The larger the area, the more accurate the diagnostic test is. AUROC can be measured by the following equation:

$$AUC = \int_0^1 ROC(t)dt$$

where t = (1 – specificity) and ROC (t) is sensitivity.

In short, AUROC curve is a good tool to select possible optimal cut-point for a given diagnostic test.

8.3.4 Sensitivity, Specificity and Accuracy

This section will focus on sensitivity, specificity and accuracy in the context of patients' re-admission. Sensitivity is a related concept, which measures the percentage of people who experienced a re-admission and are correctly identified by the model as being at risk. Specificity is defined as the proportion of people who did not experience a re-admission and were correctly identified as being at low risk. Calculation of sensitivity, specificity and accuracy is explained below with the help of table 7.

	Predicted Output		
Observed Output		Negative	Positive
	Negative	a (True Negative)	b (False Positive)
	Positive	c (False Negative)	d (True Positive)

Table 7 Conditions of terms used in the discrimination measurements.

Table 8 shows that the total number of patients (T) as given by a + b + c + d. Discrimination measures of our proposed model can be described as below. The number of patients who actually did not have a re-admission was a + b and the

number who actually had a re-admission was c + d. The number of patients who were predicted not to have a re-admission was a + c and the number predicted to have a re-admission was b + d. The following five measures are used in this project to show how accurately the models predict re-admission.

- The **percentage accuracy in classification** is the percentage of patients that are correctly predicted as to whether or not they had a re-admission and is given by the formula (a + d) / T.
- The **sensitivity** of the model is the percentage of patients that actually had a re-admission that were correctly predicted and is given by d / (c + d).
- The **specificity** of the model is the percentage of patients that actually did not have a re-admission that were correctly predicted and is given by a / (a + b).
- The **positive predictive value** is the percentage of patients predicted to have a re-admission that were predicted correctly and is given by d / (b + d).
- The **negative predictive value** is the percentage of patients predicted not to have a re-admission that were predicted correctly and is given by a / (a + c).

Measure	Description	Calculation
Accuracy	Probability to correctly Classify outcome.	$\frac{a}{a+b+c+d}$
Sensitivity	Probability to predict positive outcome when true state is positive.	$\frac{a}{c+d}$
Specificity	Probability to predict negative outcome when true state is Negative.	$\frac{a}{a+b}$
Positive Precision	Probability to correctly classify outcome predicted to be positive.	$\frac{d}{b+d}$
Negative Precision	Probability to correctly classify outcome predicted to be negative.	$\frac{a}{a+c}$

Table 8 Discrimination measures for Fuzzy regression Model.

The prediction performances of the proposed model and the traditional model are described based on the values of sensitivity, specificity and accuracy described above in Table 7 and Table 8.

8.4 Validation Exercise

This section contains an overview of the predictive modelling techniques used in our research. Three techniques of logistic regression, classification trees and neural networks were compared to predict the likelihood of re-admission. The outcome is obtained between 0 and 1, which is the scenario of re-admission. The value is then multiplied by 100 to give the risk score of re-admission. A risk score closer to 100 represents a very high risk of re-admission. An individual with a risk score of 50 or above are predicted to have a high risk of re-admission. Patients with a risk score of less than 50 are predicted to have low risk of re-admission.

All three techniques use values of one or more independent variables (either binary or continuous in nature) to predict whether a patient had a re-admission (the dependent binary variable). This section describes the three techniques.

8.4.1 Fuzzy Regression Model Validation

The model validation is done based on the framework developed in chapter 6. Our proposed model adapts fuzzy regression method which models uncertain relationship between health variables and risk of re-hospitalization of patients. Following (Tanaka & Watada, 1989), our proposed model included a fuzzy output, and a non-fuzzy input vector. Our fuzzy regression algorithm has used the linguistic term “risk of re-admission”. Crisp output “re-admission” is fuzzified using the membership function. This process is known as fuzzification. A membership function is used to quantify a linguistic term “risk of re-admission” into “high, medium and low“risk of re-admission.We have used triangular and trapezoidal membership function. Fuzzy regression method models the relationship between significant independent variables and response variable “risk of re-admission”.

8.4.1.1 Significant independent variables included in the fuzzy regression model

The derived model uses a small set of variables which includes:

- Patient's age
- Whether current admission was an emergency admission (defined in HES as an admimeth from 21-28)
- Whether there had been an admission in the past 6 months and 12 months.
- History of admissions in the prior five years (from prior HES diagnostic field).
- Charlson comorbidity severity index (CCSI) used in calculation of total severity index score.
- Reference conditions (Reference_condition) calculated using HRG codes

8.4.1.2 Dependent Variable

As shown in chapte4, dependent variable "risk of re-admission" is a fuzzy variable. Fuzzy variables are used for representing imprecise numerical quantities in a fuzzy environment. Risk of re-admission can be represented by linguistic variable. A linguistic variable is generally decomposed into a set of linguistic terms of "high, medium or low" risk of re-admission. Membership functions are used in the fuzzification of crisp output to fuzzy linguistic terms. The transition from high risk of re-admission to low risk of re-admission can be shown by gradual transition from high to low, which is shown by fuzzy set with degree of membership in figure 19 and 20. The form of membership function such as triangular, trapezoidal membership function used in our research.

The dataset was partitioned into two sections as all the predictive models are constructed or built on **training dataset** and then the performance of the model is validated or tested using a **validation dataset**. Therefore, the full dataset of 109,245 rows was randomly split using a random selection so that 60% of the rows (65,547) were used for the training dataset and the remaining 40% of rows (43,698) were used for the validation dataset.

Unusual or extreme observations (outliers) for interval or continuous (non-binary) variables are usually removed from training datasets prior to the application of

predictive algorithms to ensure that the models are built using stable and consistent data. Therefore, the extreme top and bottom 0.1% of values for the interval variables were removed from the training dataset. This accounted for 655 rows and left 64,892 rows remaining in the dataset. Outliers are not removed from the validation dataset as each of the models should be tested on actual data to determine their true performance. The performance of the model is tested on the validation set.

The model correctly classified 72.5% of the 64,892 in the training dataset correctly as to whether the patient had an emergency readmission within 12 months. The model correctly classified 72.5% of the 64,892 cases in the training dataset correctly as to whether the patient had an emergency readmission within 12 months. This is known as the percentage accuracy and the corresponding value from the validation data set was 71.6%.

For our model validation, we have divided our dataset into training and validation dataset. Of the 64,892 training set records 23,245 actually had a re-admission and the remaining 41,647 did not. The sensitivity of the model is the percentage of records that actually had a re-admission within 12 months that were correctly predicted to have a re-admission by the model (known as true positives). For a risk threshold of 50, of the 23,245 training dataset patients that actually had a re-admission within 12 months, 58.8% (13,668) were correctly predicted as having re-admission.

The specificity of the model is the percentage of records that actually did not have a re-admission within 12 months that were correctly predicted not to have a re-admission by the model (these are also known as **true negatives.** Of the 41,647 training dataset patients that actually did not have a re-admission within 12 months, 87.4% (36,339) were correctly predicted as not having a re-admission.

The **positive predictive value is** the percentage of records that the model predicts will have a re-admission that actually did have the re-admission. The positive predictive value for the training dataset of the analysis showed that 72.02% (13,668) of the 18,976 patients predicted to have a re-admission actually did. The positive predictive value for fuzzy regression method is better than logistic regression

method. **The negative predictive value** shows that 79.1 %(36,339) of the 45,916 patients predicted not to have a re-admission within 12 months were predicted correctly.

Figure 32 shows the percentage of patients flagged by the algorithm as being likely to have a re-admission that actually went onto have the re-admission. The horizontal axis shows the risk score threshold and this refers to the cut off level by which a person is predicted as having a re-admission. Figure 32 shows the effect that varying the risk score threshold level has on the percentage of flagged patients who were readmitted. Higher risk score threshold result in higher percentages of flagged patients actually having re-admissions. The model used in this thesis (the blue and red lines) for percentage of flagged patients in 6 months and 12 months, and green line for percentage of flagged patients in 30 days. At risk score threshold of 50, the percentage of flagged patients who were readmitted within 12 months appears to be better than the percentage of flagged patients in 6 months and 30 days. The reason that percentage of patients flagged by 12 months is better that readmission within 30 days because less than one tenth of patients are readmitted within 30 days of discharge.

As shown in Figure 32 percentage of patients flagged within 6 months and 12 months appears at risk score threshold of 85 appears to be similar.

Our model tested on validation set, and it gives almost similar results as in training set as shown in figure 33. Comparison of our model result is done with PARR1 model also, which is shown in detail in Appendix 5. As compared with PARR model our model gives better result in terms of percentage of flagged patients.

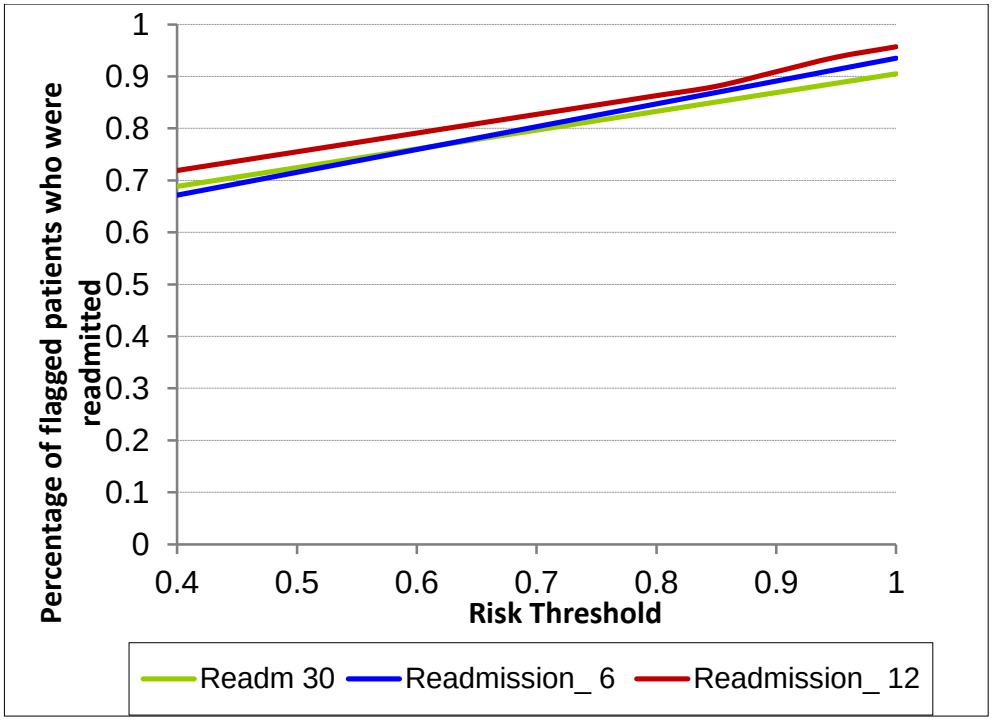

Figure 32 The percentage of patients flagged by the fuzzy regression model

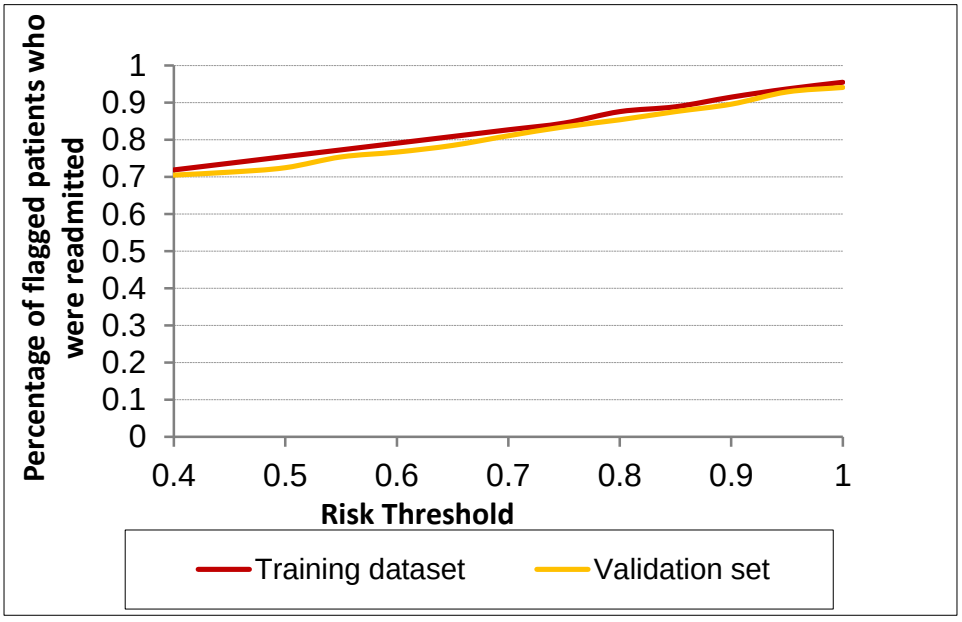

Figure 33 The percentage of patients flagged by the fuzzy regression model by training and validation set

8.4.1.3 Conclusion of Fuzzy Regression Model.

The findings are similar and compared with other models which are explained in later sections with high numbers of previous admissions being a strong predictor of re-admission. As described in later sections, all three classification tree, logistic regression models and fuzzy regression models are similar in ability to extract factors that are significant in predicting re-admission. Fuzzy regression model found less independent variables that are significant in predicting re-admission as opposed to 19 found by logistic regression models. Details of significant independent variables identified by logistic regression is given in section 8.4.2.1 The variables found to be significant in the tree model were also found to be significant in the fuzzy regression model proving that these were highly significant in both models. The factor of age, which is highly significant in the fuzzy regression but is not significant in classification tree model, returns to be significant again in the logistic regression model. However, the factor of age still plays a more significant role in the fuzzy regression model. In the next section, we have compared our model with traditional methods (logistic regression, neural network, and decision tree).

8.4.2 Logistic Regression

Logistic regression predicts the probability of the outcome (re-admission) occurring given actual values of the independent variables. The general form for the logistic regression equation showing the probability of the outcome occurring is given by equation below

$$P(R) = \frac{1}{1 + e^{-(\beta_0 + \sum_n^1 \beta_n X_n)}} \qquad (1.21)$$

The terms used in the above equation are as follows

P(R) is the probability of re-admission

R is the outcome of re-admission

e is the nature logarithm base

β_0 is the intercept or constant term in the regression equation

β_n are the coefficients (weightings) for the n independent variables used to predict the dependent variable

X_n are the n independent variables used to predict the dependent variable

The parameters above are all determined by fitting a model (with the independent variables that are most helpful in predicting re-admission) to the observed data so that the error between the actual observed outcomes and predicted outcomes are minimised.

A **stepwise** logistic regression procedure is used when assessing which independent variables are important (or are significant) in predicting the dependent variable. The stepwise model works by initially trying to include all the independent variables to predict the dependent variable. Then a process begins whereby all the independent variables that are above the significance threshold of 0.05 (i.e. those that are unimportant in predicting re-admission) are excluded one by one. This process continues until all that remain are significant predictors. Each time a variable is excluded the variables that have previously been excluded are retested in case they warrant re-inclusion into the model. All independent variables which have a significance value of less than 5% in adding to the predictive ability of the model are used. The performance of the model is then validated using the validation dataset.

The **Wald** statistic is used in logistic regression to identify independent variables that are significant predictors in a model. It tells us whether the β_n coefficient for the nth independent variable is significantly different from zero. If it is significantly different from zero then the independent variable adds to the predictive ability of the model and it has a significant value of less than 0.05.

Exp(β_n) in logistic regression is the change in odds of the outcome (re-admission) occurring given a unit change in the nth independent variable with all other independent variables being controlled for. For binary independent variables such as gender, the odds ratio tells us how much more likely a patient is to have a re-

admission if they are male compared to female while all other factors are controlled. For continuous variables such as the average number of episodes per spell in the previous 3 years, the odds ratio tells us the increased chances of re-admission for a one unit increase in the number of episodes per spell. More details of logistic regression can be found in section 2.1 of Appendix 2.

8.4.2.1 Significant independent variables included in the Logistic regression model

The following output (Table 6) shows the independent variables that were found to be significant (at the 5% level) in predicting re-admission within 12 months. These were the only independent variables which added significantly to the predictive power of the model.

Parameter	Estimate (β)	Standard Error	Wald	Significance	Exp(β)
Age 75 plus at admission (0)	-0.2222	0.0116	366.43	<.0001	0.641
Average number of episodes per emergency admission spell	0.1936	0.0152	162.09	<.0001	1.214
Average number of episodes per non-emergency admission spell	0.1270	0.0206	38.15	<.0001	1.135
Number of emergency admissions within the previous 5 years	0.1999	0.00746	717.67	<.0001	1.221
Number of emergency admissions within the previous 6 months	0.2071	0.0152	186.57	<.0001	1.230
Number of non-emergency admissions within the previous 5 years	0.0274	0.00423	42.01	<.0001	1.028
Reference condition in the previous 5 years (0)	-0.1267	0.0131	93.67	<.0001	0.776
Respiratory Infection (0)	-0.1005	0.0173	33.62	<.0001	0.818
Severity Index	0.0709	0.00954	55.28	<.0001	1.074
White (0)	-0.0856	0.00956	80.24	<.0001	0.843
Alcohol (0)	-0.2095	0.0245	73.13	<.0001	0.658
Cancer (0)	-0.0822	0.0204	16.18	<.0001	0.848
CTDRA (0)	-0.1458	0.0345	17.91	<.0001	0.747
Development disability (0)	-0.1701	0.0570	8.92	0.0028	0.712
Diabetes (0)	-0.0873	0.0176	24.64	<.0001	0.840
Drug abuse (0)	-0.2283	0.0442	26.63	<.0001	0.633
Injury from fall (0)	0.1278	0.0159	64.99	<.0001	1.291
Ischaemic heart disease (0)	-0.0320	0.0155	4.26	0.0390	0.938

Table 9 Significant Independent variables included in logistic regression model.

The estimate (β) column in Table 11 shows the coefficients that are used in the model to predict whether a record will fall into the category of having a re-admission or not. Positive β values mean that an increase in that independent variable will lead to an increase in the chances of re-admission. Negative β values mean that an increase in that independent variable will lead to a decrease in the chances of re-admission. For example, as the coefficient for the variable 'average number of episodes per emergency admission spell' is positive (0.1936) then an increase in this variable will mean an increase in the chances of re-admission. For another example, as the variable 'age 75 plus at admission **(0)**' is negative (-0.2222) then the nearer to zero this value is then the smaller the chances are of a re-admission.

The column titled Exp (β) represents the odds ratios for each of the independent variables. For binary independent variables the odds ratio tells us the chances of someone having a re-admission when being in one group compared to the other. The table 12 shows that if a patient is aged under 75 on admission then the chances of re-admission are reduced by a factor of 0.64 compared to those aged 75 or over, with all other factors being controlled for. Another example is that those not admitted due to injury from a fall are 1.29 times more likely to have a re-admission than those admitted for this reason, with all other factors being controlled for.

For continuous variables such as the number of emergency admissions within the previous 5 years the odds ratio is 1.22 and means that for each extra emergency admission that the patient has, the odds of a re-admission increases by 1.22, while all other factors are being controlled.

The independent variables which increase the chances of a re-admission are if the patient is in one or more of the following groups:

- Age 75 plus at admission
- Are of White ethnic origin
- Having a high total severity index score
- Having a high average number of episodes per emergency admission spell

- Having a high average number of episodes per non-emergency admission spell
- Having a high number of emergency admissions within the previous 5 years
- Having a high number of emergency admissions within the previous 6 months
- Having a high number of non-emergency admissions within the previous 5 years
- Having any of the following
    - Respiratory Infection
    - Alcohol abuse
    - Cancer
    - Connective tissue disease/rheumatoid arthritis (CTDRA)
    - Development disability
    - Diabetes
    - Drug abuse
    - Ischaemic heart disease
- An admission for a reference condition
- Having an admission for some other reason than an injury from a fall

The Wald value is an indication of how important the independent variable is in the predictive ability of the model. The higher the Wald value the more influence the variable has in the model. The four variables with the largest Wald values and therefore the greatest ability to predict the dependent variable are:

- Number of emergency admissions within the previous 5 years (717.67)
- Age 75 plus at admission (366.43)
- Number of emergency admissions within the previous 6 months (186.57)
- Average number of episodes per emergency admission spell (162.09)

8.4.2.2 Logistic Regression Model Performance

The method of logistic regression uses a stepwise process whereby the most useful independent variables in terms of predicting the dependent variable are included in the model and the insignificant predictors are left unused. All independent variables which have a significance value of less than 5% in adding to the predictive ability of the model are used.

Of the 64,892 training set records, 23,245 actually had a re-admission and the remaining 41,647 did not. The sensitivity of the model is the percentage of records that actually had a re-admission within 12 months that were correctly predicted to have a re-admission by the model (known as true positives). Of the 23,245 training dataset patients that actually had a re-admission within 12 months, 41.2% (9,576) were correctly predicted as having re-admission.

The specificity of the model is the percentage of records that actually did not have a re-admission within 12 months that were correctly predicted not to have a re-admission by the model (these are also known as **true negatives**). Of the 41,647 training dataset patients that actually did not have a re-admission within 12 months, 90.3% (37,607) were correctly predicted as not having a re-admission.

The **positive predictive value is** the percentage of records that the model predicts will have a re-admission that actually did have the re-admission. The positive predictive value for the training dataset of the analysis showed that 70.3% (9,576) of the 13,616 patients predicted to have a re-admission actually did, which is not as better as fuzzy regression which with positive predictive value of 72.02% (13,668)

The **negative predictive value** is the percentage of records that the model predicts will not have a re-admission that actually did not have the re-admission. The negative predictive value for the training dataset of this analysis showed that 73.3% (37,607) of the 51,276 patients predicted not to have a re-admission were predicted correctly.

Fig 41 shows the percentage of patients flagged by the algorithm as being likely to have a re-admission that actually went on to have the re-admission. The horizontal axis shows the risk score threshold and this refers to the cut off level by which a person is predicted as having a re-admission. The output from logistic regression gives us the percentage chance that a person will have a re-admission. Fig 41 shows the effect that varying the risk score threshold level has on the percentage of flagged patients who were re-admitted. Higher risk score threshold result in higher percentages of flagged patients actually having re-admissions. The model used in this thesis (the blue and red lines) for percentage of flagged patients in 6 months

and 12 months, and green line for percentage of flagged patients in 30 days. The percentage of flagged patients who were readmitted within 12 months appears to be better than the percentage of flagged patients in 6 months and 30 days. As shown in figure 41, percentage of patients flagged within 12 months and 30 days at risk score threshold of 40 appears to be quite similar, whereas percentage of patients flagged within 6 months is not as better as readmission in 12 months.

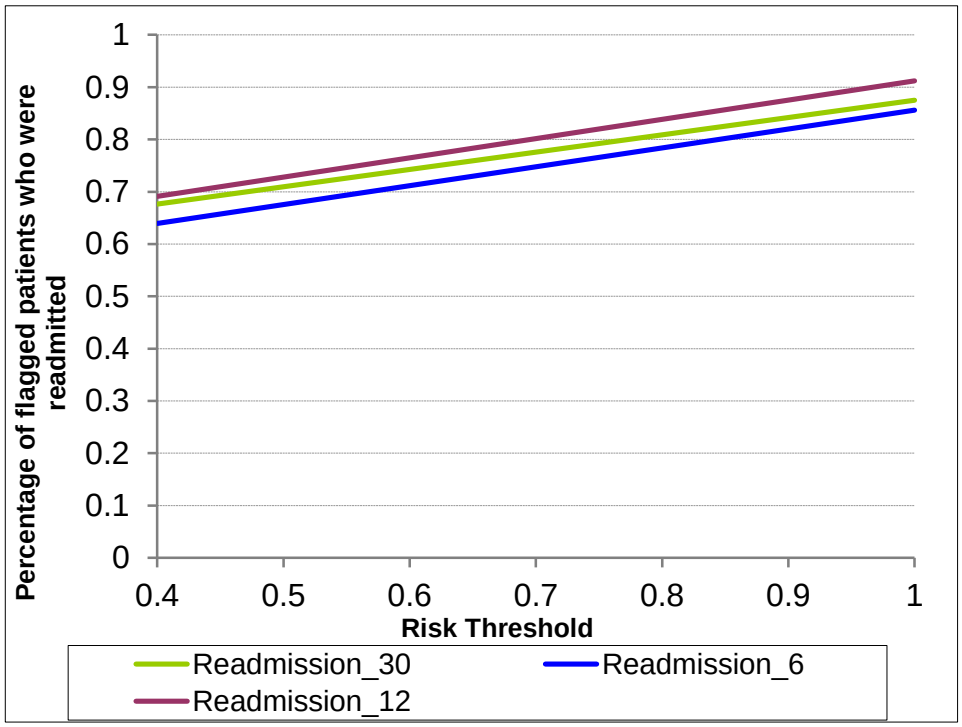

Figure 34 Percentage of patients flagged by the logistic regression model

8.4.3 Decision Tree

Classification tree model predicts one attribute (the dependent variable) given other attributes (the independent variables). Classification is the process of assigning a discrete value (or class such as whether the patient has a re-admission) as accurately as possible to an unlabelled and previously unseen record. Each time we receive an answer to a question we can ask a further question until we can be fairly confident that a patient will or will not have a re-admission. For example, we might first ask if the patient is aged 75 or over. If the answer is yes, then we may ask another question enquiring if the patient is male. If the answer is again yes, the next question might enquire about the number of times that the patient has been admitted in the previous 5 years. If this value is over 5 then the model might give

us a risk score of say 90% that the patient will have a re-admission in the next 12 months. We can therefore formalise rules that predict a patient's chances of re-admission. This type of model is known as a classification tree (Razi & Kuriakose, 2005; Austin, 2007) because patients are given risk scores of re-admission by the answers to several layered questions. These layers can be viewed visually as having a tree structure with

- **a root node** by which the data is initially split
- **internal nodes** representing questions to which the data is interrogated and split by. The higher up the tree a question is asked, the more likely it is to play a decisive role in predicting re-admission.
- **branches** representing a split from a question node (e.g. under 75 years of age or 75 and over), and
- **terminal nodes (leaves or leaf nodes)** which define an output class that allows us to classify a patient by their risk of re-admission.

More details of theoretical study can be found in section 3.2 of Appendix 3.Our focus is on the classification tree model which was fitted to the dataset. The same training and validation datasets which were used in the logistic regression analysis were used here for model performance comparison purposes. Therefore, the input dependent and independent variables used in the classification tree model were the same as those used in the logistic regression model.

Tree algorithm used	C4.5
Maximum number of branches from a node	4
Maximum depth of tree	4
Minimum number of observations in a leaf	250
Observations required for a split search	600

Table 10 Settings used for the classification tree model

8.4.3.1 Significant independent variables included in the model

The classification tree shows the following significant factors in table 8 predicting re-admission (the variable names as shown in the tree are displayed along with the relative importance in model.

Factor	Factor name in tree	Relative importance in model
The number of emergency admissions within the previous 5 years	*NumberOfEMAD_within_5years*	1.000
The severity index total score for conditions in the current admission and in the previous 5 years	*Severity_Index*	0.246
The number of emergency admissions within the previous 6 months	*NumberOfEMAD_within_6months*	0.068
Whether the patient had a reference condition in the current admission or in the previous 5 years	*Ref_condition_prev_5_yrs*	0.060

Table 11 Settings used for the classification tree model.

All of the above factors were also significant in the logistic regression model. The number of emergency admissions in the previous 5 years is the single most important predictor to the model, having a relative importance which is 4 times higher than the next most important factor (severity index). It is interesting to note that factors such as age and ethnic origin which were significant in predicting re-admission in the logistic regression model are not significant in this model. However, the general characteristics for patients who are readmitted or not are summarised below.

8.4.3.2 Factors that increase the chances of re-admission

The number of emergency admissions in the previous 5 years is the most dominant factor in predicting a re-admission, with larger numbers of previous admissions greatly increasing the chances of a return to hospital. Number of previous re-admissions act as a strong and accurate predictor. Patients with 4 or more previous admissions will have a re-admission, and with 6 or more previous admissions being a particularly strong and accurate predictor. Recent history is important in predicting the chances of re-admission as the likelihood of a return to hospital increases if the patient has had 5 emergency admissions in the previous 5 years and at least 1 of these was in the previous 6 months. Re-admission chances are increased if the patient has had severe conditions or many conditions in the previous 5 years and presence of a reference condition also increase the chances of re-admission.

8.4.3.3 Classification tree model performance

Both classification tree and logistic regression models are similar in able to extract factors that are significant in predicting re-admission. Classification tree found only five independent variables that are significant in predicting re-admission as opposed to 19 found by logistic regression models. The 5 variables found to be significant in the tree model were also found to be significant in the regression model proving that these were highly significant in both models.

This analysis focuses on the classification tree model which was fitted to the dataset. The same training and validation datasets which were used in the logistic regression and fuzzy regression analysis were used here for model performance comparison purposes. Therefore the input dependent and independent variables used in the classification tree model were the same as those used in the logistic regression model.

The **sensitivity** of the model shows that 43.7% (10,158) of the 23,245 patients that actually had a re-admission within 12 months were correctly predicted to have the re-admission. The **specificity** of the model shows that 89.5% (37,274) of the 41,647 patients that actually did not have a re-admission within 12 months were correctly predicted not to have the re-admission.

The **positive predictive value** shows that 69.9% (10,159) of the 14,531 patients predicted to have an emergency re-admission within 12 months actually did so. This value is not quite as good as that obtained from the logistic regression 70.3%. The **negative predictive value** shows that 74.01% (37,274) of the 50,361 patients predicted not to have a re-admission within 12 months were predicted correctly.

Fig 35 shows the effect that varying the risk score threshold level has on the percentage of flagged patients who were readmitted. Higher risk score threshold result in higher percentages of flagged patients actually having re-admissions. The model used in this thesis (the blue and red lines) for percentage of flagged patients in 6 months and 12 months, and green line for percentage of flagged patients in 30 days. The percentage of flagged patients who were readmitted within 12 months appears to be better than the percentage of flagged patients in 30 days and 6 months. As shown in figure 42 percentage of patients flagged within 30 days, 6 months and 12 months appears to be similar at risk score threshold of 40. With increasing threshold between 50 to 70 risk score threshold, percentage of readmitted patients flagged within 30 days and 6 months appears to be similar, while percentage of patients flagged for readmission within 12 months is still better than readmission_6 or readmission_30.

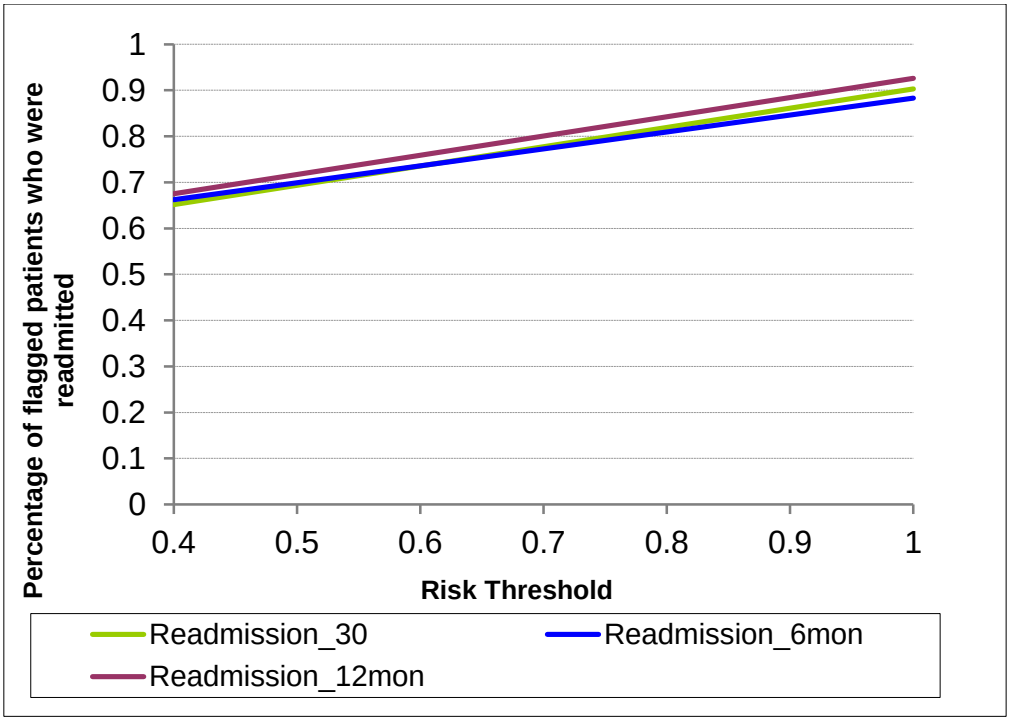

Figure 35 The percentage of patients flagged by the classification tree model

8.4.4 Neural Network Validation

Neural networks (DI-Russo, 2002; Lin, et al., 2010)are dynamic nonlinear models that are used to predict the value of a dependent variable given several independent variables. When predicting the dependent variable the model also gives the probability of obtaining the outcome (which would be re-admission in this scenario).

Neural networks are very good models for finding patterns between the dependent and independent variables by learning from the dataset and displaying the relationship between the variables. The multiple independent variables (in this project, the characteristics and hospitalisation history of the patient) all exist separately in individual neurons (nodes or cells) in the **input layer** of the neural network. The dependent variable (the binary variable showing if a patient had a re-admission or not) exists in the **output layer** of the network. Between the input and output layers there exists at least one (usually one or two) hidden layers. The input and output layers are connected by synapses (arcs) which join the input cells within the input layer to nodes within the hidden layer(s), which in turn connect to the output layer as a combination of merged factors. These arcs have **weights** which

enable us to determine the significance that the independent variable has on predicting the dependent variable. When the dataset values for the independent variables are used to train the neural network model the weights are optimised in order to give the best fitting model which is most accurate in classifying new records correctly as to whether or not they had a re-admission. Details of theoretical study of neural network model is found in section 2.3 of Appendix 3.

It focuses on the neural network model which was fitted to the dataset. Neural networks with 1 hidden layer with between 2 and 25 neurons in each layer were constructed and the settings which gave the best performance in terms of positive predictive value and percentage accuracy in classification within the training and validation datasets are shown in Table 9

Number of hidden layers	1
Number of hidden neurons	9
Network architecture	Multilayer Perceptron

Table 12 Settings used for the neural network model.

8.4.4.1 Significant independent variables included in the neural network model

Figure 36 shows blue and red boxes which represent the sign and size of the weighting from each of the 9 neurons in the hidden layer within the neural network to the output variable (re-admission within 12 months). Blue boxes (H11, H13, H14, H16 and H18) represent positive weightings with the dependent variable and red boxes (H12, H15, H17 and H19) represent negative weightings. The size of the box reflects the magnitude of the weight with boxes H17 and H14 having the largest impact on the dependent variable. The actual weightings are shown in Table 10.

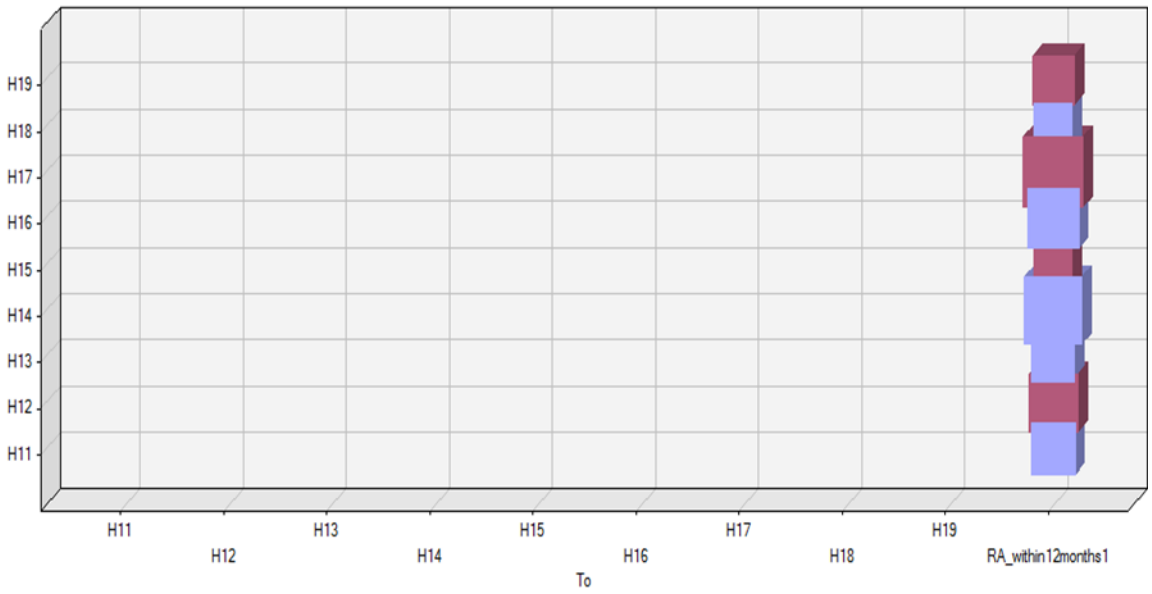

Figure 36 Blocks representing weightings on arcs between neurons.

Neuron	Weighting	Neuron	Weighting	Neuron	Weighting
H11	0.439	H14	0.757	H17	-0.831
H12	-0.566	H15	-0.318	H18	0.305
H13	0.436	H16	0.617	H19	-0.389

Table 13 Weightings on arcs between neurons.

The blue box at nodes H11, H13, H14, H16 and H18 all represent positive weights to the dependent variable, which means that positive input values to these hidden nodes increase the chances of re-admission as the positive input when multiplied by the positive output weight results in a positive (increased) chance of re-admission, while negative input values decrease the chances of a return to hospital as the negative input multiplied by the positive output weight results in a negative (decreased) chance of re-admission. Therefore any of the 26 independent variables that have positive weights to these 5 boxes increase the chances of re-admission while those with negative weights decrease the chances of re-admission. As hidden neuron H14 has the highest positive weight with the dependent variable we shall consider the most important independent variable inputs to this neuron and see

which of the variables make re-admission more and less likely. The 5 independent variables with the largest absolute weights going to node H14 are shown in Table 16. Any variable with a positive weighting increases the chances of re-admission while those with negative weightings decrease the chances of re-admission.

Variable	**Weighting**	**Absolute weighting**
Severity Index	1.006	1.006
Number of non-emergency admissions within the previous 5 years	0.583	0.583
Number of emergency admissions within the previous 5 years	0.556	0.556
Renal Failure **(0)**	0.506	0.506
Age 75 plus at admission **(0)**	-0.490	0.490

Table 14 Variables with the largest absolute weights going to node H14.

Therefore, the following factors increase the chances of re-admission

- Having a high severity index
- Having higher numbers of emergency and non-emergency admissions in the previous 5 years.
- Not having renal failure
- Being aged 75 or over on admission

Most of these factors were also found to be significant in the other two models.

7.4.4.2 Performance of Neural Network Model

The **sensitivity** of the model at a risk threshold of 50 shows that 40.01% (9,330) of the 23,245 patients that actually had a re-admission within 12 months were correctly predicted to have the re-admission. The **specificity** of the model shows that 89.9% (37,440) of the 41,647 patients that actually did not have a re-admission within 12 months were correctly predicted not to have the re-admission.

The **positive predictive value** shows that 69% (9,228) of the 13,537 patients predicted to have an emergency re-admission within 12 months actually did so. This value is not quite as good as that obtained from the logistic regression and fuzzy regression analysis at 72.02%, and 70.3%. The **negative predictive value** shows that 72.9% (37,440) of the 51355 patients predicted not to have a re-admission within 12 months were predicted correctly.

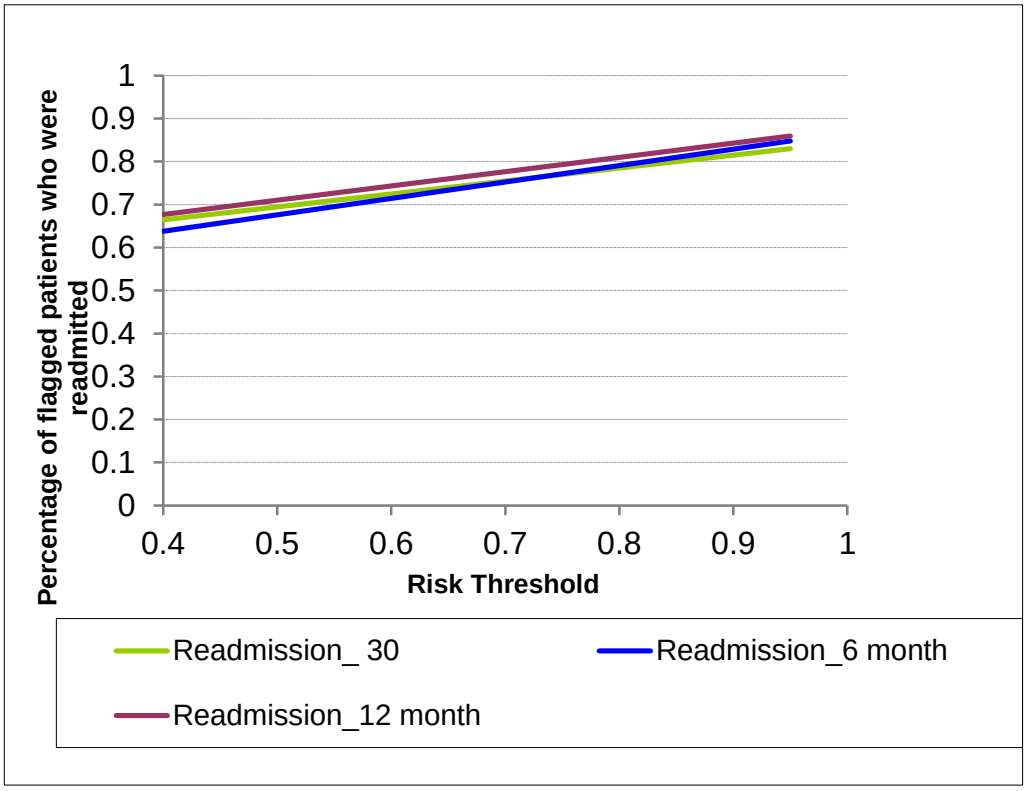

Figure 37 The percentage of patients flagged by the neural network model

Figure 37 shows the effect that varying the risk score threshold level has on the percentage of flagged patients who were readmitted. Higher risk score threshold result in higher percentages of flagged patients actually having re-admissions. The model used in this thesis (the blue and red lines) for percentage of flagged patients in 6 months and 12 months, and green line for percentage of flagged patients in 30 days. The percentage of patients flagged in 12 months at risk score threshold of 40 appears to be better than patients admitted in 6 months and 30 days. As shown in figure 44 at risk score of 60 or above percentage of patients flagged within 30 days months and 6 months appears to be similar. At higher risk score of 90 or above, readmission within 30 days, 6 month and 12 months is quite similar.

8.4.4.3 Conclusion of Neural network

The findings are similar to those of found in the previous two models with high numbers of previous admissions being a strong predictor of re-admission. The factor of age, which was highly significant in the logistic regression but was not significant in classification tree model, returns to be significant again in the neural network model. However, the factor of age still plays a more significant role in the regression model than in the neural network model.

8.4.5 Comparison of different Models

A number of predictive models and tools have also been developed for the prediction of patients who are at high risk of re-admission (Austin, 2007; Rosma, et al., 2008; Krumholz, et al., 1997). These studies tend to produce conflicting results where factors associated with unplanned re-admissions vary widely in statistical significance and, as a consequence, the predictive model and the tool may not provide sufficiently accurate predictions. Most predictive models have focused on regression techniques, although there is an emerging interest in machine learning algorithm.

Several studies have tried to compare the performance of previous predictive models. A number of studies have compared the predictive ability of decision trees with regression analysis and Artificial Neural Network. Predictive ability can be compared using the area under ROC.

The results show that fuzzy regression models provide better prediction in comparison to logistic regression models, decision tree and neural network model where the data is binary or categorical. Logistic Regression Models show better results than Neural Network. This is represented by area under ROC curve in figure 38.

The performance of the model is shown with the help of ROC curve in the figure 38. The receiver operating curve in the figure 38 illustrates the trade-offs for users between sensitivity (true positives) and 1-specificity (false negatives) for the algorithm. True positives (sensitivity) and false positives (1-specificity) are evaluated at different risk score (0-100). At risk score 50 sensitivity and specificity

are evaluated. Sensitivity of the model is the percentage of records that actually had re-admission as per the model, which is 58.8% for fuzzy regression algorithm. The specificity of the model is the percentage of records that actually did not have re-admission within 1, 6 and 12 months that were correctly predicted not to have re-admission as per the model (true negatives). For fuzzy regression model, the specificity is 87.4%. A table summarizing the sensitivity, specificity and PPV for different models is given here:

Method	**Sensitivity**	**Specificity**	**PPV**
Fuzzy Regression	58.8%	87.4%	72.02%
Logistic Regression	41.2%	90.3%	70.3%
Decision Tree	43.7%	89.5%	69.9%
Neural Network	40.01%	89.9%	69%

Table 15 : Summary the sensitivity, specificity and PPV for different models.

The figure 38 shows that fuzzy regression model performs well with addition of significant independent variables. Area under curve (AUC) is 0.735 which is slightly higher than logistic regression (AUC 0.723), and better than decision tree (AUC 0.715). Neural network shows the least performance with AUC 0.699. Area under ROC curves for different models with confidence interval values is given below:

Method	**AUC**	**Confidence Interval**
Fuzzy Regression	0.735	95% CI:73.85%-89.88%
Logistic Regression	0.723	95% CI:72.43%-86.54%
Decision Tree	0.715	95% CI:71.56%-87.24%
Neural Network	0.699	95% CI:69.83%-87.54%

Table 16 Area under ROC curves for different models with confidence interval values.

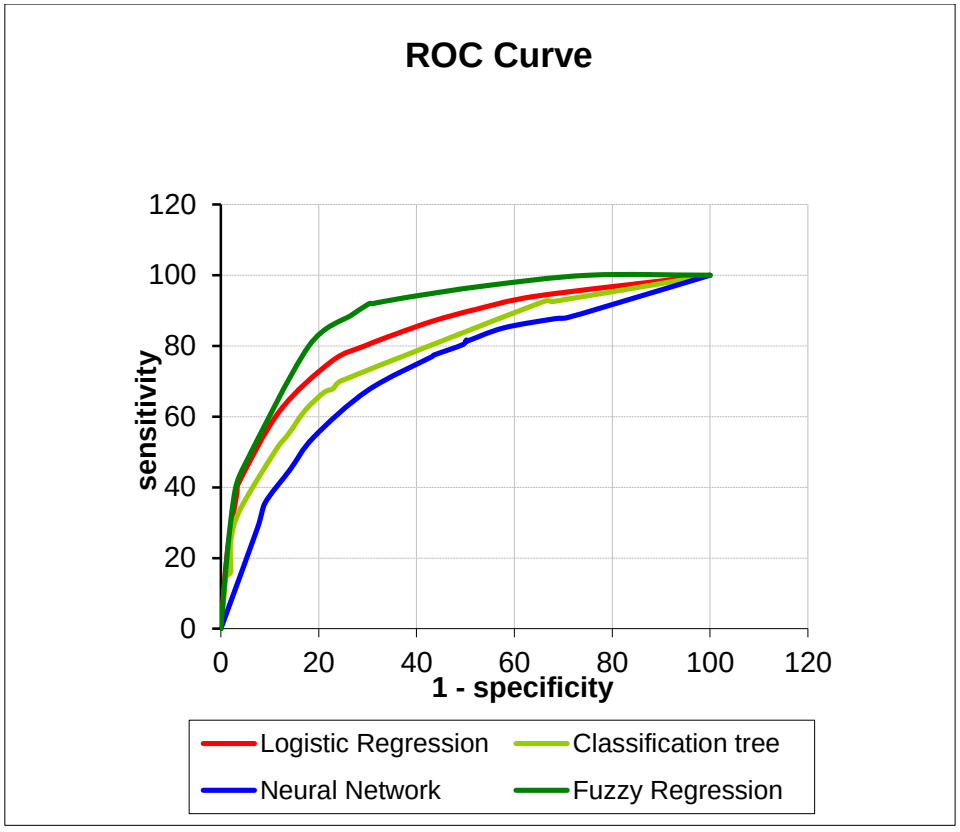

Figure 38 Comparison of ROC curves for different models

Figure 39 shows the percentage of patients flagged by the algorithm as being likely to have a re-admission that actually went onto have the re-admission. The horizontal axis shows the risk score threshold and this refers to the cut off level by which a person is predicted as having a re-admission. Figure 46 shows the effect that varying the risk score threshold level has on the percentage of flagged patients who were readmitted. Higher risk score thresholds result in higher percentages of flagged patients actually having re-admissions. The fuzzy regression performed better than other models such as logistic regression, decision tree and neural network model in our experiments. At risk score of 40 or above, the percentage of patients flagged by fuzzy regression model appears to be better than logistic regression method, decision tree, and neural network. Although at risk score of 40 and 50, percentage of flagged patients by neural network appears to be quite similar to decision tree, is not as good as fuzzy regression models. The percentage of patients flagged as readmitted by logistic regression and classification tree model is almost similar. In conclusion, neural network model does not perform as good as other models.

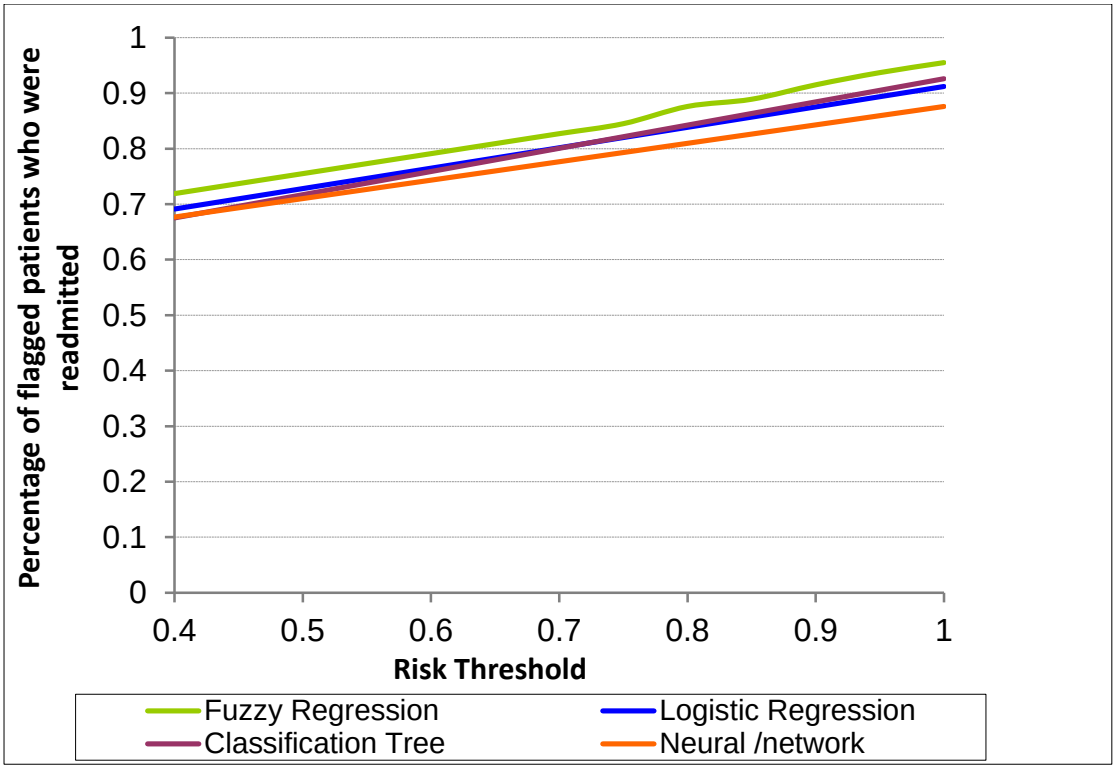

Figure 39 Comparison of the percentage of patients flagged by the neural network model.

8.5 Benefits of Predictive Models

There are a wide range of predictive models available to the NHS in England for forecasting health and social care outcomes. Most of these models aim to predict readmissions within time frame, however they differ in terms of time period they predict and whether they predict single or multiple readmissions. For example, models predict readmissions within 12 months of discharge but models could also be developed for shorter and longer time frames. In order to compare different models, it is important that their predictions are based on the same source of healthcare data. The term predictive model or risk prediction model are used interchangeably to implement the algorithm or the computational steps to evaluate individuals at risk of readmission. We have implemented various algorithms such as logistic regression, decision tree and neural network to make predictions about their readmission. These models forecast patients at risk of readmission using risk scores. Technically, PARR and PARR+ models are usually logistic regression models, but decision tree and neural network models are

developed and implemented for comparison and evaluation purpose. Our model implements fuzzy regression algorithm to estimate individuals at likelihood of readmission.

Predictive risk models can be useful in predicting occurrence of readmission that following benefits to the healthcare community.

- By predicting such events, it is possible to provide preventive care that can help in improving health status and quality of care to the high risk individuals.
- By providing preventive care, it can aid in general net savings after taking into account the success rate of interventions and its cost.
- Predictive risk models are developed by analyzing historical data and correlations between outcome of an event and a range of predictor (explanatory variables) from prior data. Significant predictor variables may include age, gender, and a range of diagnostic variables. Clinicians may benefit from understanding a range of predictor variables responsible for outcome of an event, and in targeting individuals for whom intervention may reduce or delay hospital readmission.

8.6 Summary

In this chapter, different risk prediction models are compared. These models tend to produce results with significant factors associated with re-admissions. All of the methods tend to find factors that are significant in predicting re-admission. The method of logistic regression uses a stepwise process whereby the most useful independent variables in terms of predicting the dependent variable are included in the model and the insignificant predictors are left out. All of the factors significant in the logistic regression model were also significant in classification tree model. Classification tree, neural network and fuzzy regression method found less significant independent variables as opposed to 19 factors found in logistic regression. The factor of age, which was highly significant in the logistic regression but was not significant in classification tree model, returns to be significant again in the fuzzy regression model. Two of the factors which were significant in all models are previous admissions and severity of illness. Addition of significant

independent variables may increase the predictive power of the model. These models were also compared for area under ROC curves, where fuzzy regression method (with AUROC as 0.735 fuzzy) outperforms other methods.

Chapter 9

9. Conclusions and Future work

9.1 Conclusion

The overall aim of this research was to develop a framework that can be used for the prediction of risk of re-admission of a patient within 12 month of discharge. Risk of readmission can be stratified into high, medium or low risk of readmission. Boundaries of risk stratification is not crisp, and proper intervention can aid in movement of patients from high risk to medium or low risk band. Readmission of a patient can be defined in a set as [0, 1], where re-admitted is '1" and not-readmitted is "0". Uncertainty in decision making can be defined as a degree of opinion of a readmitted patient belonging to a set [0, 1]. Traditional techniques are not able to deal with uncertain type of information. In our case, the response variable is governed by uncertain relationship among response variable and explanatory variables. A framework adapting fuzzy linear regression method was proposed to provide solution to the problem described.

We will revisit research objectives and verify whether they have been successfully fulfilled. Basically there were three objectives: development of a framework and a model that predicts patients at risk of re-admission, development of novel algorithm that captures uncertainty in "risk of re-admission", and performance evaluation of the developed model. In the process, we also identified significant independent variables for the proposed model. This framework was developed to deal with uncertainty in risk of re-admission. The overview of the framework developed to achieve these objectives was described in chapter 5 of the thesis whereby a full description of steps followed was provided. In chapter 5, we have described the development of proposed novel algorithm which adapts fuzzy regression method. This algorithm was based on fuzzy linear regression method with inclusion of significant risk factors. Independent variables were assessed with the help of statistical analysis. More detailed description of health system variables

and significant risk factor for identifying patients at risk of re-admission is given in chapter 4.

The proposed fuzzy regression model was developed as a machine learning algorithm that can be used to identify patients at risk of re-admission. This algorithm has been developed and validated on Hospital Episode Statistics (HES) dataset and validated using validation dataset. The validation exercise was performed by constructing other models like decision tree, artificial neural network, and logistic regression models that were also tested on the same dataset. The model performances were measured by area under receiver operating curve (ROC). The explanatory power of the model, which refers to its ability to explain certain dependencies in the data, was investigated using statistical analyses. The percentage of patients flagged to have re-admission within 12 months is predicted using positive predictive value, at varying risk threshold levels.

The findings show that:

- The fuzzy regression model, artificial neural network and logistic regression model exhibit significant similar prediction ability based on similar dataset extracted from HES data.
- The fuzzy regression, neural network and logistic regression prediction models exhibit better prediction ability when significant input variables are included in the model.
- The proposed fuzzy regression model has an advantage over artificial neural network because of its ability to explain the associations between the predictor variables and predicted outcome. The discussion on associations is explained in chapter 5 and chapter 7.
- Both, the fuzzy regression model and logistic regression model are able to interpret the associations between input predictor variables and response variables. However, logistic regression model is not able to handle uncertain response variable, and uncertain relationship between response variable and predictor variables. We overcame this limitation in our fuzzy regression based proposed algorithm, as it handles uncertainty in risk of re-admission.

- Our proposed prediction model is better than classical prediction model when prediction performance is considered. Prediction performance is evaluated and compared in chapter 7.The proposed framework using novel algorithm is highly recommended because of its good performance in predicting patients at risk of re-admission. It also has the ability to capture uncertain nature of re-admission and handle uncertain response variable.
- We used risk scores ranging from (0-100) for patients readmitted within 12 months discharge period using HES data. Patients can be stratified into various risk bands from high to low according to these risk scores. This would help us in identification of percentage of patients flagged by the model at various risk threshold levels. The performance of predictive model could be compared and evaluated with the help of area under receiver operating curve, specificity and sensitivity at a risk score threshold value. We captured and compared results of our proposed model with other traditional models. By risk stratification, interventions could be targeted for individuals who are most in need of hospital resources.

The outcome of this research contributes towards a better understanding of risk of re-admission.

As a conclusion, the proposed fuzzy regression model is highly recommended to be used as an aid to predict patients at risk of re-admission.

The extensive experiment and validation exercises carried out for this research work has led to the comparison with other risk prediction models. The information may serve as the guide for future prediction modelling as well as providing foundation for setting up of a computer based prediction tool.

9.2 Limitation of Studies

The research mainly concentrated on design and development of a framework to predict patients at risk of re-admission using HES dataset. This framework is a step towards handling uncertain information. However, there are a number of issues

regarding the dataset and a proper way to handle uncertain information. We have tried to address some of the limitations of our research here:

- The research has relied on historical data from HES (Hospital Episode Statistics) data for past five years from 1999/2000 to 2004/2005 for triggering admission in 2004/2005. There were concerns for missing episodes, which affected the results of the predictive model. Also, the data collected has quality and consistency issues. The data was missing date of birth and discharge dates for the patients.
- Outlier detection problem: As for graphical analysis there is a lack of tools that address the outlier problem in fuzzy framework. When outliers exist in the data, the interval obtained by using fuzzy linear regression or logistic regression becomes too large and thus result in erroneous power.
- In our framework we have modelled non-crisp output as “risk of re-admission” which can be represented by linguistic variable “high, medium or low” risk of re-admission. We have modelled response variable with membership functions as triangular and trapezoidal membership functions. Significant input variables could also be modelled with membership functions. It is difficult to represent all input variables with membership functions in a fuzzy environment. Consideration of membership functions of input variables could increase computational complexity of our algorithm.
- In our research, we have adapted fuzzy regression method to estimate unknown dependency between risk of re-admission and input variables. The dependency between response variable “risk of re-admission” and independent variables can be masked by different fuzzy rules. In addition, relationship between input variables is unknown, therefore significant number of fuzzy rules could be generated. A large of number of fuzzy rules would make our algorithm computationally intensive and slow.

  This is explained with the help of small example. The investigation on health system variables is done with a single input variable fed into the system and the output was recorded or measured. Eventually, the number of input variables fed into the system can be increased to two, three and four etc. Each input predictor variable can be represented by a fuzzy membership function. For a two input variable

system, where fuzzy relationship between variables is considered a number of fuzzy rules will be generated. In the first case, if we consider fuzzy rules for two input variables out of all input variables (for e.g 8 input variables), the number of possible combinations of input set will be $^{8}c_{2}$, which will be 28 input sets to consider. Following, if 3-variables are considered with fuzzy membership function, then the total number of input sets to be experimented will be 56 ($^{8}c_{3}$). Similarly, if 4-input variables are represented with fuzzy membership function, and fuzzy relationship between variables is considered there will be total of 70 ($^{8}c_{4}$) input fuzzy sets. Increasing the number of input variables with fuzzy membership functions will increase possible number of input sets. Using all possible input sets will require extensive computational time and expensive resources.

- We have considered uncertainty in the form of fuzziness in "response variable". But there could be other types of uncertainties present in variables which we have not considered. As for example, uncertainty may be in the form of randomness. These other types of uncertainties may be considered simultaneously along with fuzziness.

9.3 Novel elements of Research

Our research is novel in the sense that it is the first study that handles linguistic uncertainty between high, medium and low risk of readmission.Traditional methods of prediction such as logistic regression are not able to account for uncertain nature of risk of hospital re-admissions.

The proposed model helps to account for the uncertain nature of risk of re-admission. It also allows to stratify patients into different risk bands and targeting interventions at individuals who will benefit most. The model can be used to set risk scores and thresholds for patients who are at risk of future re-admissions.

Patients at risk of re-admission could be identified with consideration of significant variables. Selection of independent variables as potential covariates helps in recognizing important risk factors to predict likelihood of re-admission. This proposed model based on significant risk factors selected by the novel algorithm predicts patients at high risk of re-admission. The knowledge of the impact of risk

factors will provide clinicians better decision-making and reducing the number of patients re-admitted to the hospital.

The other important novel element is the design and development of an algorithm that adapts fuzzy regression method. This algorithm estimates the unknown dependency between the independent variables and the response variable. We believe that this will start a novel approach to handle uncertain nature of outcome of an event using the possibilistic approach. This model could be a new approach to modeling uncertainty in risk of re-admission, and can help in better decision making.

9.4 Future Work

We plan to apply the fuzzy linear regression method with more datasets by including more significant health system variables for predicting patients at risk of admission and stratifying them from high to low risk of re-admission. We wish to visualize the results to check the impact of these significant risk factors in risk of admission of a patient. We can further investigate the risk of re-admission with input variables using statistical techniques. Response variable may be influenced by one or multiple risk factors, therefore statistical methods such as multivariate analysis could be used.

The relationship between the response variable and risk factors may not always be linear. Therefore, we can propose another algorithm adapting fuzzy logistic regression method with non-linear to linear transformations. The fuzzy regression method could be further improved with least square estimates. Fuzzy least square estimate can be used to estimate the errors, and unknown dependency between risk of admission and explanatory variables.

Based on proposed framework, we plan to develop a prototype adapting our novel algorithm for predicting patients at risk of re-admission. A predictive tool can be designed with the use of HES datasets for stratification of patients into high, medium and low risk of re-admission. By using risk prediction tool, it would be possible to identify those patients who are most in need of hospital resources and stratify them according to complexity of utilisation of resources.

This model for identifying high risk patients will enable GPs and Clinical Commissioning Groups (CCGs), to target specific groups of patients and enable clinicians to offer better preventive care for high risk individuals. On the basis of proposed model, a tool could be developed to predict such high risk events. It will be important for local NHS organizations to consider the potential role of these tools as improving the health of their population. It will also be important in identifying individuals who can be benefit by tests and treatment. Our approach is useful for CCGs in moving patients from high to medium, or medium to low risk by offering interventions to avoid readmissions

Appendix 1

Authors(s)	Objective	Methodology
(García-Pérez et al., 2011; Billings et al., 2013)	Identify risk factors for hospital re-admission.	Logistic Regression
(Chan et al., 2011)	Identifies admission and unplanned re-admission of COPD patients	Univariate analyses, includes t-tests and chi-square tests. Multivariate analyses, uses logistic regression.
(Demir et al., 2008)	To determine the risk of unplanned re-admission.	Phase-type distribution and transition modelling framework.
(Marcantonio et al., 1999)	Identify high risk of re-admission.	A case-control design, the Cochrane Mantel-Haenszel chi-square to assess bivariable associations and conditional logistic regression model to determine independent associations with re-admission.
(Brunetto, A. T. et al, 2010)	To identify the pattern and risk of unplanned hospital admissions in a dedicated phase I clinical trials unit.	Logistic Regression model was applied to define the baseline characteristics associated with the unplanned admissions.
(Byrne et al., 2010)	To examine in-hospital mortality and its predictors in all elderly patients	Admission case mix system for Elderly (ACME) that uses multivariate logistic regression model by adjusting the univariate predictors of outcome.
(Corrigan and Martin, 1992)	To identify multiple hospital admissions, relationship of patient and health system characteristics associated with re-admission.	Predictive Model using regression model to predict likelihood of re-admission.

Table 1 Studies for predicting unplanned admission.

Risk factors for admission of a patient	Response variable	Authors
Age	Independent variable	(Brunetto, A. T. et al, 2010 ; Chan , et al., 2011; Corri gan and Martin, 1992; Marcantonio et al., 1999)
Severity of illness	Independent variable	(Brunetto, A. T. et al, 2010)
Type of care	Independent variable	(Pearson et al., 2002)
Morbidity/comorbidity	Independent variable	(Marcantonio et al., 1999; Bisserier et al., 2010)
Functional disability	Independent variable	(Bisserier et al., 2010)
Prior admission	Independent variable/Response variable	(Bisserier et al., 2010 ; Marcantonio et al., 1999)

Table 2 Risk Factors for Risk of Re-admission.

Appendix 2

A2.1 Hospital Episode Statistics (HES) Data base: Conditions used in this Thesis

For each of the conditions in TableA1.1 binary variables are created to record whether each of the patients involved in 109,243 triggering emergency admissions had the condition in the current admission in in the previous five years.

If the patient had the condition during this period then they were allocated a value of 1 for the condition. For example, if a patient had cancer at one point in either the triggering admission or in the previous 5 years then they were coded as having cancer = 1. It was also recorded as to whether each patient had any of the reference conditions shown in Table A1.3. If the patient had any of the reference conditions the variable of reference_condition was set to 1, else it was 0.

The Charlson comorbidity severity index allocates severity scores to diseases (as shown in Table A1.2). This was incorporated into this analysis by summing up the total severity of conditions that the patient had in the triggering admission and in the 5 years prior to that point. The following formula was used in Access to work out the severity index for patients. All the condition variables are binary in nature. Therefore, if the patient had the condition the corresponding variable is set to 1. The sum of all conditions at the same severity weighting is calculated and then multiplied by the weighting to produce the severity total score. Patients with higher total severity index scores either had more severe conditions (such as HIV or Metastatic Cancer) or just had multiple conditions.

Severity_Index =
(([IschaemicHD]+[CHF]+[PVD]+[CVD]+[Mental]+[COPD]+[CTDRA]+[Peptic_Ulcer]+[Liver]+[Diabetes_without_comps])*1)+(([Hemiplegia]+[RenalFail]+[Diabetes_with_comps]+[Cancer_lower_form])*2)+([Mod_Sev_Liver]*3)+(([Metastatic_Cancer]+[HIV])*6)

Additional numerical variables were created to record the number of previous emergency and non-emergency admissions that the patients had in the 1 month

(30 days), 6 months (180 days), and 12 months (365) prior to their triggering admission date. The average number of episodes per emergency admission and non-emergency admission spells in the 5 years prior to their triggering admission were also created for these patients. It was important here to exclude all duplicate episodes for the patients. Therefore previous admissions that started on the same day were only counted once.

Condition	**ICD 10 codes**
Alcohol abuse	F10, K70
Anaemia	D50, D64, D539, D639
Angina	I20
Asthma	J45-J46
Atrial fibrillation	I471, I48
Cancer	All codes beginning with C, D00-D48
Cerebrovascular disease (CVD)	I60-I67, I69, G45, H340, R298, R470
Congenital disability	Q00-Q99
Congestive heart failure (CHF)	I50, I110, I130
Connective tissue disease/rheumatoid arthritis (CTDRA)	M32-M36, M05, M06, M08, I39, I528, I418, I328, J990, G737
Chronic obstructive pulmonary disease (COPD)	J43-J44
Development disabilities	F70-F89
Diabetes	E08-E14, G632, H360, H280, O24
Drug abuse	F11-F16, F18-F19, K71
HIV/AIDS	B20-B24
Hypertension	I10-I15

Injury from fall	W00-W19
Ischaemic heart disease	I21-I25
Mild Liver disease	K703, K743-K746, K760, K769
Mental illness	F00-F09, F17-F69, F90-F99
Peripheral vascular disease (PVD)	I700-I702, I71-I72, I731-I739, I709, I792, I771, R2
Renal Failure	N18-N20, Z940
Sickle cell disease	D57
Respiratory infection	J02, J93, J85, J81, J32, J90, J86, J96, J393, R091, R098, J869, J998, J840-J841, J04-J06, J20-J21, J384-J387

Table A2.2 – Conditions and their associated ICD 10 codes

The condition categories of Ischaemic heart disease and mental illness in Table A2.2 above were added in place of Myocardial Infarct and Dementia, which were included in the actual Charlson comorbidity severity index measure. These two added categories were deemed to be similar in nature to the original conditions and were used instead to be consistent with the conditions used in Table A2.1

Condition	Charlson Comorbidity Severity Index	ICD 10 codes
Ischaemic heart disease	1	I21-I25
Congestive heart failure (CHF)	1	I50, I110, I130
Peripheral vascular disease (PVD)	1	I700-I702, I71-I72, I731-I739, I709, I792, I771, R2
Cerebrovascular disease (CVD)	1	I60-I67, I69, G45, H340, R298, R470
Mental illness	1	F00-F09, F17-F69, F90-F99
Chronic obstructive pulmonary disease (COPD)	1	J43-J44
Connective tissue disease/rheumatoid arthritis (CTDRA)	1	M32-M36, M05, M06, M08, I39, I528, I418, I328, J990, G737
Peptic Ulcer	1	K25-K28
Mild Liver Disease	1	K703, K743-K746, K760, K769
Diabetes without complications	1	E100, E10l, E106, E108, E109, E110, E111, E116, E118, E119, E120, E121, El26, E128, El29, E130, E131, E136, E138, E139, E140, E141, E146, E148, E149
Hemiplegia	2	G041, G114, G801, G802, G81, G82, G830-G834, G839
Renal Failure	2	N18-N20, Z940

Diabetes with complications	2	E102-E105, E107, E112, E115, E117, E122-E125, E127, E132-E135, E137, E142-E145, E147
Cancer	2	All codes beginning with C, D00-D48
Moderate to severe Liver Disease	3	I850, I859, I864, I982, K704, K711, K721, K729, K765, K766, K767
Metastatic Cancer	6	C77-C80
HIV/AIDS	6	B20-B24

Table A1.2 – Conditions with Charlson comorbidity severity index weightings and ICD 10 codes

In Table A1.3 references are made to >69, >49, <70 etc. These refer to the patient's age. Therefore as an example, 'Chronic Pancreatic Disease <70' means patients that have chronic pancreatic disease that are under 70 years of age. There are also references to w cc and w/o cc. These stand for with complications and without complications. Therefore, as another example, 'Epilepsy >69 or w cc' refers to patients with Epilepsy who were either over 69 years of age or had complications.

Condition	HRG code
Multiple Sclerosis or other CNS Demyelinating Cond	A18
Epilepsy >69 or w cc	A29
Bronchiectasis	D16
Cystic Fibrosis	D17
Chronic Obstructive Pulmonary Disease or Bronchiti	D20
Asthma >49 or w cc	D21
Fibrosis or Pneumoconiosis	D26
Other Respiratory Diagnoses >69 or w cc	D33
Complex Elderly with a Respiratory System Primary	D99
Heart Failure or Shock >69 or w cc	E18
Heart Failure or Shock <70 w/o cc	E19
Coronary Atherosclerosis >69 or w cc	E22
Arrhythmia or Conduction Disorders >69 or w cc	E29
Angina >69 or w cc	E33
Complex Elderly with a Cardiac Primary Diagnosis	E99
Large Intestinal Disorders >69 or w cc	F36
Chronic Pancreatic Disease <70	G25
Inflammatory Spine, Joint or Connective Tissue Dis	H25
Skin Ulcers	J38
Diabetes with Hypoglycaemic Emergency >69 or w cc	K11
Diabetes with Hyperglycaemic Emergency >69 or w cc	K13
Diabetes with Lower Limb Complications	K17
Complex Elderly with an Endocrine or Metabolic Sys	K99
Kidney or Urinary Tract Infections >69 or w cc	L09
Cystic Fibrosis	P02
Blood Cell Disorders	P23
Cardiac Conditions	P25
Peripheral Vascular Disease >69 or w cc	Q17
Coagulation Disorders	S04
Red Blood Cell Disorders >69 or w cc	S05
Red Blood Cell Disorders <70 w/o cc	S06
Senile Dementia	T01

Table A1.3 – Reference conditions and their HRG code

Appendix 3

3.1 Logistic Regression

Logistic regression predicts the probability of the outcome (re-admission) occurring given actual values of the independent variables. The general form for the logistic regression equation showing the probability of the outcome occurring is given by equation A2.1 below

$$P(R) = \frac{1}{1 + e^{-(\beta_0 + \sum_n^1 \beta_n X_n)}} \qquad (\boldsymbol{A3.1})$$

The terms used in the above equation are as follows

- P(R) is the probability of re-admission
- R is the outcome of re-admission
- e is the nature logarithm base
- β_0 is the intercept or constant term in the regression equation
- β_n are the coefficients (weightings) for the n independent variables used to predict the dependent variable
- X_n are the n independent variables used to predict the dependent variable
- The parameters above are all determined by fitting a model (with the independent variables that are most helpful in predicting re-admission) to the observed data so that the error between the actual observed outcomes and predicted outcomes are minimised.

The **– 2 log likelihood (-2LL)** value in logistic regression tells us how much information remains unexplained (in terms of predicting re-admission) after the model has been fitted. Therefore, it is a measure of how well the model actually works. If after fitting the model the -2LL value is large then the model does not fit the data very well as there are large amounts of re-admissions that could not be predicted correctly. However, if the -2LL is small then the model works well. The formula for -2LL is given by equation A2.2 below

$$-2LL = -2\sum_{i=1}^{n}\Big(R_i \ln P(R_i) + (1 - R_i)\ln\big(1 - P(R_i)\big)\Big) \qquad (\mathbf{A3.2})$$

Where

- R_i is the actual observed outcome (i.e. re-admission or not) for the ith person
- $P(R_i)$ is the predicted probability that re-admission occurs for the ith person
- ln is the natural logarithm
- n is the number of observations (patients) in the sample

To assess the success of the fitted model its -2LL figure is compared to a **baseline (intercept only) -2LL** figure. The -2LL baseline figure represents a very basic model in which the only value used to predict re-admission is the intercept **(β_0)** term in equation A2.2. Therefore, as all the β_n values are 0 in the baseline model no independent variables are used to predict re-admission.

The value of -2LL obtained when using the more sophisticated model with the inclusion of the independent variables is called the **new (intercepts and covariates) -2LL** figure. Therefore the improvement (in terms of how much the new sophisticated model adds in explaining re-admission over the baseline model) is given by the **Chi-square likelihood ratio** shown by equation A2.3 below

$$\chi^2 = \big(-2LL(baseline)\big) - \big(-2LL(new)\big) \qquad (\boldsymbol{A3.3})$$

If this value is large enough then the addition of the independent variables has added to the predictive power of the model. In which case, the model will have a **significance value** of less than 0.05.

There are three separate measures that are used in logistic regression to assess how well the model fits the data in terms of the variation that can be explained in the dependent variable by the independent variables. The first measure is the

Hosmer and Lemeshow R square (R_L^2) figure, which is given by equation A2.4 below

$$R_L^2 = \frac{Chi\ square\ likelihood\ ratio\ (\chi^2)}{-2LL(baseline)} \qquad (\boldsymbol{A3.4})$$

This measures the proportion of the unexplained information in the baseline model that was explained by the new model and it varies in value between 0 and 1. A value close to 1 indicates that the new model was extremely good at being able to explain the variation in re-admission across patients in the dataset.

The second measure is the **Cox and Snell R square (R_{CS}^2)** value which is given by equation A2.5 below

$$R_{CS}^2 = 1 - e^{\left(\frac{1}{n}\left((-2LL(new))-(-2LL(baseline))\right)\right)} \qquad (\boldsymbol{A3.5})$$

The third measure is the **Nagelkerke R square (R_N^2)** value which is given by equation A2.6

$$R_N^2 = \frac{R_{CS}^2}{1 - e^{-\left(\frac{(-2LL(baseline))}{n}\right)}} \qquad (\boldsymbol{A3.6})$$

In equations A2.5 and A2.6 the n stands for the number of patients in the sample and e is the nature logarithm base.

The Cox and Snell R square and Nagelkerke R square values are usually similar in value to the Hosmer and Lemeshow R square value and they all give a measure as to the predictive power of a model.

-2 Log Likelihood		Likelihood ratio Chi-Square	Degrees of freedom	Significance
Intercept only	Intercept and covariates			
83486.044	72116.839	11369.2051	19	<.0001

Table A3.1 Likelihood Ratio Test for Global Null Hypothesis: BETA=0

Table A2.1 shows the significance level for the model performance is 0.0001 which means that the addition of the independent variables has added to the predictive power of the model. Therefore the model with the independent variables included is better than our naïve model where all observations were assigned to the group of no re-admission. The **Hosmer and Lemeshow** value of 0.14 suggests that the inclusion of the independent variables has added some predictive ability to the model. The value for Hosmer and Lemeshow varies between 0 and 1 with 0 meaning that the variables have added no predictive ability whatsoever. However, as our figure is significantly different from zero we can conclude that the inclusion of the additional variables has helped in this model.

The **Cox and Snell R square** and **Nagelkerke R square** figures give us two slightly different measures for the amounts of variation which is explained in the dependent variable (re-admission within 12 months) by the independent variables. The two figures for this analysis are 0.16 and 0.22 respectively indicting that between 16% and 22% of the variation in re-admission within 12 months within the 69,342 training set records can be accounted for by the significant independent variables.

2.2 Classification trees

Classification is the process of assigning a discrete label value (or class such as whether the patient has a re-admission) as accurately as possible to an unlabeled and previously unseen record. This is achieved by a classification tree model, which

predicts one attribute (the dependent variable) given other attributes (the independent variables).

The classification trees used in this thesis are mainly constructed using the **C4.5 algorithm** which is also known as **Entropy reduction**. The steps involved in the C4.5 algorithm are shown below

Step 1: Calculate the **purity** or **amount of information** we would require to correctly classify a class (re-admission) of a new instance (patient) using the initial dataset. This is known as **entropy** (**information** or **info**) and is a measure of how pure the initial dataset is. The entropy value ranges from 0 to 1 and if all the instances (observations) in the initial dataset belong to one class the entropy or additional information needed to specify the class of a new instance would be 0. If half the instances belong to class 1 and the other half belong to class 2 in the initial dataset then a great deal of extra information would be required to correctly classify a new instance based on the initial dataset (in this situation the entropy would be equal to 1). The following equation is used to calculate the initial purity of the dataset D in respect to the binary dependent variable of re-admission

$$Info(D) = -\sum_{i=1}^{2}\big((freq(Class_i, D)/|D|) \times log_2(freq(Class_i, D)/|D|)\big) \qquad (\boldsymbol{A3.7})$$

$Class_i$ refers to each of the class outcomes of the dependent variable (in this example these would be re-admission and non-re-admission), $freq(Class_i,D)$ is the number of times that class i occurs in the training dataset and |D| refers to the number of rows in the dataset. If for example a dataset contained 100 patients, of which 80 did not have a re-admission (NR) and 20 did have a re-admission (R) then |D| = 100, freq(NR,D) = 80 and freq(R,D) = 20.

Once the information value is calculated for the initial dataset we have a baseline model which the performance of further more sophisticated models can be compared against. If the information value for the initial dataset is near 1 then we

would hope that the more sophisticated models (using the independent variables) would reduce this figure.

Step 2: Test each attribute (independent variable) to determine which one is the best to define as the root node and to initially split the data by. This is achieved by selecting each attribute in turn and placing it at the root of the tree and determining the average entropy (pureness) value at child nodes immediately following the initial split. The best attribute to split on will be the one that causes the highest gain in information from the initial information value. (i.e. the greatest gain corresponds to the greatest reduction in information needed to correctly classify the class attribute). Gain in information is calculated as the initial entropy value before the data were split minus the new average value. Therefore we split by the attribute which creates the purest child nodes and highest information gain for the model.

If the independent variable is binary then there are only two possible branches from a node when splitting on that variable. If the independent variable is continuous then threshold values at which to split the data are determined. For example, the number of admissions in the previous 5 years is a continuous variable which may range from say 0 to 10. This variable would be examined to find the best cut off values for predicting re-admission and then the observations would be split into two or more groups (e.g. under 2 admissions, 2 to 5 admissions and over 5 admissions). Therefore, classification trees treat continuous data like discrete data by forming groups.

The information value of the model after splitting by each variable is recorded using the formula

$$\text{Info}_{\text{var i}}(T) = \sum_{i=1}^{n}\left((|T_i|/|T|) \times Info(T_i)\right) \qquad (\boldsymbol{A3.8})$$

Where var i is the ith independent variable used to split the data by, n is the number of child nodes formed from the split and T_i represents each subset of data following the split. The information value obtained from splitting the data by each of the

independent variables should be recorded. The independent variable which has the smallest information value should be used to split the data as this will result in the largest information gain over the baseline model.

Step 3: The process in steps 1 and 2 is repeated recursively for each child node created from the previous split, using only those instances that reach the node. If at any time all instances at a node have the same classification (i.e. entropy of 0), stop developing that part of the tree as the node is totally pure and is a leaf node. Alternatively, we stop when the data cannot be split any further, even if the leaf node is impure.

The secondary tree algorithm which is briefly used in this project to confirm the robustness of the results from the C4.5 algorithm is called CART or Gini reduction. This method is similar to the C4.5 algorithm but instead of using the entropy formula a slightly different measure (the Gini measure) is used to determine the the best variable to split the data by and the pureness of the dataset at each split.

2.3 Neural networks

Neural networks are very good models for finding patterns between the dependent and independent variables by learning from the dataset and displaying the relationship between the variables. Figure A2.1 shows a typical neural network with four independent variables in the input layer, one dependent variable in the output layer and one hidden layer consisting of two hidden nodes. This type of neural network is called a feed forward multilayered perceptron neural network as the learning takes place from the input layer to the output layer as values flow through the model. The arrows in Figure A2.1 indicate the direction of the flow of the data through the model

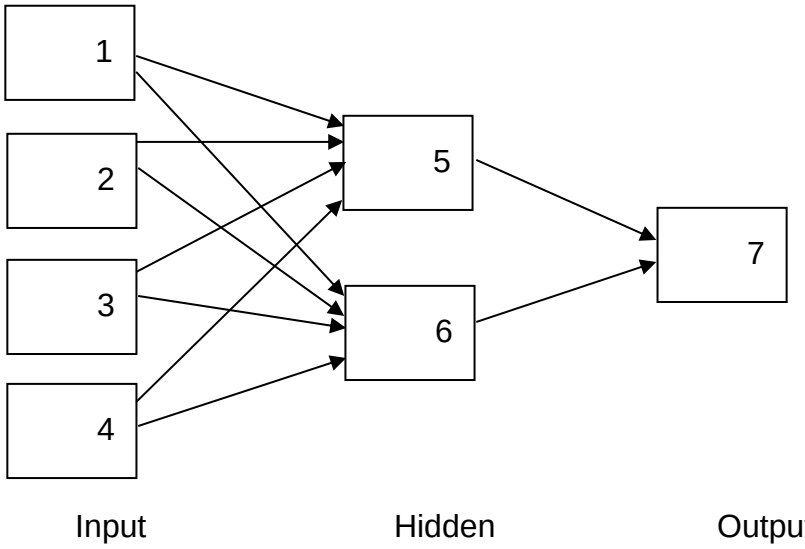

Figure A3.1 – A neural network with one hidden layer

In Figure A2.1 there are ten synapses or arcs and each of these carries a weight. Weights are defined as W_{ij} which means the weight on arc from node i to node j with i = 1, 2, 3, 4, 5 and 6 and j = 5, 6 and 7.

The input data values received through the four input nodes for the independent variables used in this model are transformed so they lie between 0 and 1. Therefore any binary input variables can remain as they are, but those that are continuous or categorical and take values greater than 1 or less than 0 are transformed to the [0,1] range. Equation A2.9 shows the **transformation function** that is used to transform the input variable values to [0,1].

$$new\ input\ value = \frac{original\ input\ value - minimum\ value\ for\ that\ variable}{range\ of\ input\ values\ for\ that\ variable} \quad (\boldsymbol{A3.9})$$

The terms in equation A2.9 are defined as follows

- new input value is the transformed input value for that variable in the interval [0,1]
- original input value is the original input value for that variable
- minimum value for that variable is the smallest value in the dataset for that variable
- the range of input values for that variable is the difference between the largest and smallest values in the dataset for that variable.

After scaling the input values for nodes 1 to 4 to lie in the range [0,1], the input to hidden nodes 5 and 6 are calculated by summing the product of each input value by the weightings on the synapses to the hidden nodes. Therefore the general formula to calculate the input to nodes 5 and 6 is shown below

$$Input\ to\ hidden\ node\ j = \sum_{i=1}^{n} Input\ value\ i * W_{ij} \qquad (\boldsymbol{A3.10})$$

The terms in equation A2.10 are defined as follows

- Input to hidden node j is the computed input value that arrives at node j
- i refers to input node number i which ranges from 1 to n. In Figure 3.1 n = 4 as there are four input nodes joined to hidden nodes 5 and 6.
- Input value i is the input value arriving at node j from node i.
- W_{ij} is the weight on the arc from node i to node j.

Once the input to each hidden node is calculated it needs to be transformed so it lies between [0, 1] before it can be passed to the next layer. The function used to transform the value is called the **sigmoid function**, which is given by equation (A2.11)

$$Output\ from\ node\ j = f(input\ to\ node\ j)$$
$$= \frac{1}{1 + e^{-input\ to\ node\ j}} \qquad (\boldsymbol{A3.11})$$

The terms in equation A2.11 are defined as follows

- Output from node j is the transformed value in the interval [0,1] that is passed to the next layer
- Input to node j is the combined input to node j from the previous layer
- e is the nature logarithm base

Using the same process defined above, the transformed output from the hidden nodes then get combined with the weights to the output layer before the final predicted output is obtained by using the sigmoid transformation function on the combined input to the output node. This output (which is in the interval [0, 1] gives us the probability of the outcome occurring (i.e. re-admission for the patient).

While the neural network model is being constructed using the dataset values the weights on the synapses which result in the minimum error between the observed and predicted outcome variable are obtained using the process of **backpropagation**. Backpropagation works by feeding the set of observations into the model for the first patient and the output value probability of re-admission is obtained. The predicted output value is then compared with the actual output value in the dataset and an error value is obtained. This output error is then propagated back through the network and all of the synapse weights are amended to reduce the error as much as possible. This process is carried out multiple times with data values for other patients until the weightings are optimal.

Neural networks are seen as having a wow factor as they mimic the human brain. The results from neural networks are often extremely good because of the nonlinear nature of the models. However, as the models are very complicated it is often difficult to comprehend the results from neural networks.

Appendix 4

Code Extracts

SQL exhibit A: Code used to obtain the full 3,500,058 last episodes of emergency admissions that started and ended in 2004/05

```
select row_ind, endage, startage, mydob, dob_cfl, ethnos, hesid, sex, admi_cfl,
admidate admission_date,
str_to_date(concat(left(admidate,2),'/',left(substring(admidate,3),2),'/',right(admidat
e,4)),'%d/%m/%Y') admidate_date,disdate discharge_date,
str_to_date(concat(left(disdate,2),'/',left(substring(disdate,3),2),'/',right(disdate,4)),'
%d/%m/%Y') disdate_date , dis_cfl, admimeth, dismeth, spelbgin, epiend, epistart,
speldur, spelend, epidur, epiorder, epie_cfl, epis_cfl, epistat, diag_01, diag_02,
diag_03, diag_04, diag_05, diag_06, diag_07, diag_08, diag_09, diag_10,
diag_11, diag_12, diag_13, diag_14, resgor, epikey, right(epikey,2) epikey_L2C,
disdest, provspno, resha, hatreat, resladst, oacode6, ward91, resro, rotreat,
hrgorig, postdist

from HES0405.data0405

where admimeth in ("21", "22", "23", "24", "28")

and admi_cfl in ("0")

and sex in ("1","2")

and dob_cfl in ("0")

and dismeth not in ("4","8","9")

and
str_to_date(concat(left(admidate,2),'/',left(substring(admidate,3),2),'/',right(admidate,4)
),'%d/%m/%Y') between '2004-04-01' and '2003-05-31';
```

SQL exhibit B: Code used to obtain a sample of 109, 243 of the full 3,500,058 last episodes of emergency admissions that started and ended in 2004/05

```
select row_ind, endage, startage, mydob, dob_cfl, ethnos, hesid, sex, admi_cfl,
admidate admission_date,
str_to_date(concat(left(admidate,2),'/',left(substring(admidate,3),2),'/',right(admidat
```

```
e,4)),'%d/%m/%Y') admidate_date,disdate discharge_date,
str_to_date(concat(left(disdate,2),'/',left(substring(disdate,3),2),'/',right(disdate,4)),'
%d/%m/%Y') disdate_date , dis_cfl, admimeth, dismeth, spelbgin, epiend, epistart,
speldur, spelend, epidur, epiorder, epie_cfl, epis_cfl, epistat, diag_01, diag_02,
diag_03, diag_04, diag_05, diag_06, diag_07, diag_08, diag_09, diag_10,
diag_11, diag_12, diag_13, diag_14, resgor, epikey, right(epikey,2) epikey_L2C,
disdest, provspno, resha, hatreat, resladst, oacode6, ward91, resro, rotreat,
hrgorig, postdist

from HES0405.data0405

where admimeth in ("21", "22", "23", "24", "28")

and admi_cfl in ("0")

and sex in ("1","2")

and dob_cfl in ("0")

and dismeth not in ("4","8","9")

and
str_to_date(concat(left(admidate,2),'/',left(substring(admidate,3),2),'/',right(admidate,4)
),'%d/%m/%Y') between '2004-04-01' and '2005-03-31'

and right(epikey,2) in ("01","31","61");
```

A4.1 Learning step: Generate an FRBS model

```
object.reg <- frbs.learn(data.train, range.data, method.type, control)

 ## Predicting step: Predict for newdata

res.test <- predict(object.reg, data.tst)

## Display the FRBS model

summary(object.reg)

## Plot the membership functions

plotMF(object.reg)


data(Sample_HES)
```

```
## Shuffle the data
## then split the data to be training and testing datasets
Sample_HES Shuffled <- Sample_HES [sample(nrow(Sample_HES)), ]
Sample_HES Shuffled[, 14] <- unclass(Sample_HES huffled[, 14])
tra. Sample_HES <- Sample_HES Shuffled[1 : 6000, ]
tst. Sample_HES <- Sample_HES Shuffled[6001 : nrow(Sample_HES Shuffled), 1 : 13]
real. Sample_HES <- matrix(Sample_HES Shuffled[106 : nrow(Sample_HES Shuffled),
14], ncol = 1)
## Define range of input data. Note that it is only for the input variables.
range.data.input <- apply(Sample_HES [, -ncol(Sample_HES)], 2, range)
method.type <- "FRBCS.W"
control <- list(num.labels = 7, type.mf = "TRAPEZOID", type.tnorm = "MIN",
type.snorm = "MAX", type.implication.func = "ZADEH")
## Learning step: Generate fuzzy model
object.cls <- frbs.learn(tra. Sample_HES, range.data.input, method.type, control)
## Predicting step: Predict newdata
res.test <- predict(object.cls, tst.Sample_HES)
## Display the FRBS model
summary(object.cls)
```

A4.2 Plot the membership functions

```
plotMF(object.cls)
data(ROCR.SampleHES)
pred <- prediction( ROCR. SampleHES $predictions, ROCR. SampleHES $labels )
pred2 <- prediction(abs(ROCR. SampleHES $predictions +
                rnorm(length(ROCR. SampleHES $predictions), 0, 0.1)),
```

```
      ROCR.s SampleHES $labels)
perf <- performance( pred, "tpr", "fpr" )
perf2 <- performance(pred2, "tpr", "fpr")
plot( perf, colorize = TRUE)
plot(perf2, add = TRUE, colorize = TRUE)
```

Confidence Interval Using R

```
library(gdata)
library(ggplot)
library(pROC)
data(Sample_HES)
png("ROC Curve_5.png")
roc(Sample_HES $outcome, Sample_HES$Predictor)
roc(outcome ~ Predictor, Sample_HES)
# Smooth ROC curve
roc(outcome ~ Predictor, Sample_HES, smooth=TRUE)
# more options, CI and plotting
roc1 <- roc(Sample_HES $outcome,
Sample_HES $Predictors, percent=TRUE,
        # arguments for auc
        partial.auc=c(100, 90), partial.auc.correct=TRUE,
        partial.auc.focus="sens",
        # arguments for ci
        ci=TRUE, boot.n=100, ci.alpha=0.9, stratified=FALSE,
        # arguments for plot
        plot=TRUE, auc.polygon=TRUE, max.auc.polygon=TRUE, grid=TRUE,
```

```
            print.auc=TRUE, show.thres=TRUE)


roc2 <- roc(Sample_HES $outcome, Sample_HES$predictor2,
              plot=TRUE, add=TRUE, percent=roc1$percent)


  ## Coordinates of the curve ##
  coords(roc1, "best", ret=c("threshold", "specificity", "1-npv"))
  coords(roc2, "local maximas", ret=c("threshold", "sens", "spec", "ppv", "npv"))
  ci(roc2)
  sens.ci <- ci.se(roc1, specificities=seq(0, 100, 5))
  sens.ci
  plot(sens.ci, type="shape", col="lightblue")
  plot(sens.ci, type="bars")
  plot(roc2, add=TRUE)
  dev.off()
```

A4.3 Function for surface viewer plot

```
  function [ output ] = FLP_3DMFplot( FuzzySet, FuzzyRules, FS1, FS2, N )
  % FLP_3DMFplot Plots the crisp output surface vs two of the fuzzy sets
  %   The function creates a surface plot of the crisp output with the fuzzyy
  %   set feature 1 (FS1) along the X axis, the fuzzy set feature 2 (FS2)
  %   along the y axis and the crisp output on the z axis. The data and
  %   labels used to generate the chart are returned from the output of the
  %   function.
  %
  % Input
```

```
% FuzzySet - a FuzzySet object generated by FLP_LoadFuzzySets
% FuzzyRules - a fuzzy rules object generated by FLP_LoadFuzzyRules
% FS1 - the number of the fuzzy set to plot on the x axis
% FS2 - the number of the fuzzy set to plot on the y axis
% N - the NxN size of the meshgrid to create for the crisp output
%
% Output
% output -  a struct object with the data and chart labels
%
% This section creates the grid of each N evenly spaced points (ESP) across
% the range of FS1 and FS2 in the FLP CrispInput format. This is
% accomplished for FS1 by appending a list of the ESP's N times while
% circularly shifting the position of the ESP's by one for each append.
% When this is concatenated with N replicates of the FS2 ESP's it
% represents the complete set of points on the surface.
X1    =    (FuzzySet.Range(FS1,1):(FuzzySet.Range(FS1,2)-FuzzySet.Range(FS1,1))/(N-
1):FuzzySet.Range(FS1,2))'; % the ESP's for FS1
I1 = (1:N)'; % the position index of the ESP's
X1C = repmat(X1,N,1); % the list to append (store) the values
I1C = repmat(I1,N,1); % the list to append (store) the indexes
for i = 1:(N-1) % shift and store the ESP values and indexes
   X1 = circshift(X1,1);
   X1C(i*N+1:i*N+N,1) = X1;
   I1 = circshift(I1,1);
   I1C(i*N+1:i*N+N,1) = I1;
```

```
end
X1 = circshift(X1,1); % shift them back into the starting position
I1 = circshift(I1,1);
% Replicate the FS2 ESP's N times
X2 = (FuzzySet.Range(FS2,1):(FuzzySet.Range(FS2,2)-FuzzySet.Range(FS2,1))/(N-
1):FuzzySet.Range(FS2,2))';
I2 = (1:N)';
X2C = repmat(X2,N,1);
I2C = repmat(I2,N,1);
NC = size(X1C,1); % the number of combinations
CrispInput = zeros(NC,FuzzySet.Count-1);
% Look thru the fuzzy sets and create the CrispInput table. For fuzzy sets
% other than FS1 or FS2, use the mean value for the set
for i = 1:FuzzySet.Count-1
  if i == FS1
     CrispInput(:,i) = X1C;
  elseif i == FS2
     CrispInput(:,i) = X2C;
  else
     CrispInput(:,i) = repmat((FuzzySet.Range(i,2)-FuzzySet.Range(i,1))/2,NC,1);
  end
end

% Calculate the crisp output based on the FLP functions
AntMemberGrades = FLP_Fuzzification(FuzzySet, CrispInput);
```

```
ConsqMemberGrades = FLP_FuzzyRuleEval(AntMemberGrades,FuzzyRules);
[CrispOutput, OutputMF, X] = FLP_DeFuzzification(ConsqMemberGrades, FuzzySet,
100);

% Map the crisp output into a mesh grid
CrispOutputGrid = zeros(N,N);
for i=1:NC
   CrispOutputGrid(I1C(i),I2C(i)) = CrispOutput(i);
end

% Store the data and labels for the crisp output plot
output = struct('X',X1, ...
            'Y',X2, ...
            'Z',CrispOutputGrid, ...
            'xlabel', FuzzySet.Set{FS1}, ...
            'ylabel',FuzzySet.Set{FS2}, ...
            'zlabel',FuzzySet.Set{end}, ...
            'title', 'Crisp Value Output Visualization');
% Plot the crisp output on a surface chart
surf(output.X, output.Y, output.Z);
xlabel(output.xlabel);
ylabel(output.ylabel);
zlabel(output.zlabel);
title(output.title);
end
```

5.2 Function for membership function

```
function [ mf ] = FLP_trapzMF( x,tparms )
% FLP_trapzMF Calculates the value of the trapezoidal membership function at
% each point x.
% Input
% x - a column vector of values
% tparms - a vector with the a, b, c, & d trapezoidal parameters
% Output
% a columnar vector with membership function values
mf = zeros(size(x,1),1); % preformat the output with zeros
a = tparms(1); % get the trapezoidal parameters
b = tparms(2);
c = tparms(3);
d = tparms(4);
% Use logical indexing to identify where each x value falls within the
% trapezoidal function
idx1 = x(:,1) > a & x(:,1) < b;
idx2 = x(:,1) >= b & x(:,1) <= c;
idx3 = x(:,1) > c & x(:,1) < d;
% Replicate the scalar parameters to vectors to facilitate a "vectorized"
% calculation of the membership function
a = repmat(a,size(x,1),1);
b = repmat(b,size(x,1),1);
c = repmat(c,size(x,1),1);
d = repmat(d,size(x,1),1);
```

```
% Calculate and assign the membership function for each range
res = (x-a) ./ (b-a); mf(idx1,1) = res(idx1,1);
mf(idx2,1) = 1;
res = (d-x) ./ (d-c); mf(idx3,1) = res(idx3,1);
end
```

A4.3 Function for fuzzification

```
function [ output ] = FLP_Fuzzification( FuzzySet, CrispInput )
% FLP_Fuzzification Calculates the fuzzy antecedent membership grades
%
% This function converts each of the crisp input values to a fuzzy
% membership grade for each item in the set. For example, if there are 10
% crisp input values and five items in a set, the resulting output is a 10
% by 5 matrix with the membership grades for each item being in a column
% and the rows being each crisp value. If there are multiple sets in the
% FuzzySet, the output will be a cell array with each cell containing the
% matrix of membership grades for that set.
%
% Input
% FuzzySet - a FuzzySet object generated by FLP_LoadFuzzySets
% CrispInput - a CrispInput object generated by FLP_LoadCrispInput
%
% Output
% output - the antecedent membership grades
%
```

```
output = cell(FuzzySet.Count-1,1); % pre-allocate the output
for i = 1:FuzzySet.Count % loop through each set
  if ~strcmp(FuzzySet.Set{i,1},'Output') % do not calculate grades for the Output set

    mfArray = zeros(size(CrispInput,1),FuzzySet.ItemCount(i,1)); % initialize the membership grade output matrix
    for j = 1:FuzzySet.ItemCount(i,1) % loop through each item in the set
      mfArray(:,j) = FLP_trapzMF(CrispInput(:,i),FuzzySet.Parms{i,1}(j,:)); % calc the membership grades for each CrispInput
    end
    output{i,1} = mfArray; % store the grades to a cell array
  end
 end
end

function [ output ] = FLP_LoadFuzzySets( file_dir )
```

A4.4 Function to load fuzzy sets from csv file of dataset

```
% FLP_LoadFuzzySets Reads Fuzzy Sets from comma delimited text file

%   This function reads in a CSV delimited text file containing the details
%   of the fuzzy set and formats the detail in the form of a structure
%   array
% Input
% file_dir - the path and filename of the CSV file
```

```
%
% Output
% output -  a FuzzySet object
%
% read the CSV file
fid = fopen(file_dir);
fid_read = textscan(fid, '%s %s %d %d %d %d','delimiter',',');
fclose(fid);
% ***** Section to create a unique list of the sets *****
input_ct = size(fid_read{1,1},1); % get the number of input records to check
sets{1,1} = fid_read{1,1}{1,1}; % add the first set label to the variable to hold the unique list of sets
for i = 2:input_ct % loop thru the remaining records
   find_set = fid_read{1,1}{i,1}; % read the next set label
   find_set_row = find(strcmp(find_set,sets)); % try to find the set label in the unique list
   if size(find_set_row,1) == 0 % if the result of the find is blank, add it to the unique list
      sets{size(sets,1)+1,1} = find_set; % add the label to the unique list
   end
end

% Confirm the Output set was listed last in the text file
if ~strcmp(sets{end,1},'Output')
   fprintf('ERROR LOADING FUZZY SETS: The input file must end with the "Output" set items\n\n');
```

```
    output = [];
    return
end

input_ct = size(sets,1); % get the number of sets to check
% pre-allocate collection statistics
items = cell(input_ct,1);
item_count = zeros(input_ct,1);
parms = cell(input_ct,1);
range = zeros(input_ct,2);

for i = 1:input_ct % loop thru the each set
    find_items = sets{i,1}; % read the set label
    find_item_rows = find(strcmp(find_items,fid_read{1,1})); % find the rows with items from the set
    if size(find_item_rows,1) > 0 % if there are items in the set, create a list of items
        items{i,1} = {fid_read{1,2}{find_item_rows,1}}'; % add the items to the list
        item_count(i,1) = size(items{i,1},1);
        a = double(fid_read{1,3}(find_item_rows,1)); % get the a parameters
        b = double(fid_read{1,4}(find_item_rows,1)); % get the b parameters
        c = double(fid_read{1,5}(find_item_rows,1)); % get the c parameters
        d = double(fid_read{1,6}(find_item_rows,1)); % get the d parameters
        p_combined = [a,b,c,d];
        parms{i,1} = p_combined;
        range(i,1) = min(parms{i,1}(:,1)); % get the minimum of the range for the set
```

```
        range(i,2) = max(parms{i,1}(:,4)); % get the maximum of the range for the set
    end
end
% store the output in a structure array
output = struct('Count',size(sets,1), ...
                'Set',{sets}, ...
                'Items',{items}, ...
                'ItemCount', {item_count}, ...
                'Parms',{parms}, ...
                'Range',{range});
end
function [ output ] = FLP_LoadCrispInput( FuzzySet, file_dir )
% FLP_LoadCrispInput Reads Crisp Input from comma delimited text file
%   This function reads in a CSV delimited text file containing the crisp
%   input values and creates a matrix with the values in a column
%   corresponding to the order of the sets in FuzzySet. The first row of
%   the CSV file should contain a header row that matches the name of the
% Input
% FuzzySet - a FuzzySet object generated by FLP_LoadFuzzySets
% file_dir - the path and filename of the CSV file
% Output
% output -  a CrispInput object
%
% The number of columns for this input is variable based on the number of
% fuzzy sets in the data (N-1 for the output). Create a format string that
```

```
% will read in the set count and create a file format specification
input_spec = '%s';
N = FuzzySet.Count - 1;
for i = 1:N-1
   input_spec = strcat(input_spec,' %s');
end
% read the CSV file
fid = fopen(file_dir);
fid_read = textscan(fid, input_spec, 'delimiter',',');
fclose(fid);
in_ct = size(fid_read{1,1},1)-1; % the number of crisp input values to be read
output = zeros(in_ct,N-1); % pre-allocate the output

for i = 1:N
   find_set_row = find(strcmp(fid_read{1,i}{1,1},FuzzySet.Set)); % find the set number
   if isempty(find_set_row) % if unable to find match return error message
      fprintf('ERROR LOADING CRISP INPUTS: Unable to find matching set for "%s" column\n\n',fid_read{1,i}{1,1});
      output = [];
      return;
   end
   output(1:in_ct,find_set_row) = str2num(char(fid_read{1,i}{2:end,1})); % read the values and place in set number column
end
end
```

A4.5 Fuzzy Regression Using R

```
library(gdata)
library(gplots)
library(frbs)
library(ROCR)
 ## Input data: Using the Sample_HES dataset
## then split the data to be training and testing datasets
 frbsData <- read.csv ("D:/ Sample_HES.csv")
 data.train <- frbsData$ Sample_HES. dt[1 : 6000, ]
 data.tst <- frbsData$ Sample_HES. dt[6000: 109423, 1 : 14]
 real.val <- matrix (frbsData$ Sample_HES dt[6000: 109423, 14], ncol = 1)
## Define interval of data
range.data <-apply(data.train, 2, range)
## Set the method and its parameters,
method.type <- "WM"
control <- list(num.labels = 15, type.mf = "TRAPEZOID", type.defuz = "WAM",
type.tnorm = "MIN", type.snorm = "MAX", type.implication.func = "ZADEH",
name = "sim-0")
```

Appendix 5

A5.1 Comparison of results with PARR

Figure A5.1 shows the percentage of patients flagged up by the algorithm as being likely to have a readmission that actually went onto have the readmission. The horizontal axis shows the risk score threshold and this refers to the cut off level by which a person is predicted as having a readmission. The output from fuzzy regression gives us the percentage chance that a person will have a readmission and by default we assign anybody with a value equal to or above 50% to the predicted group of readmission = 1 (yes) and anybody who has a score of below 50% to the predicted group of readmission = 0 (no). This risk score threshold can be altered so that anybody who has a score of equal to or more than say a 40% chance of readmission is assigned the predicted value of readmission = 1 (yes) or is assigned to the readmission = 0 (no) group if they have a risk score of less than 40%. Figure A5.1 shows the effect that varying the risk score threshold level has on the percentage of flagged patients who were readmitted. Higher risk score thresholds result in higher percentages of flagged patients actually having readmissions. The model used in this thesis (the red and yellow lines in Figure A5.1) appears to be better than that used in the 2006 paper (the green line in the figure A5.1).

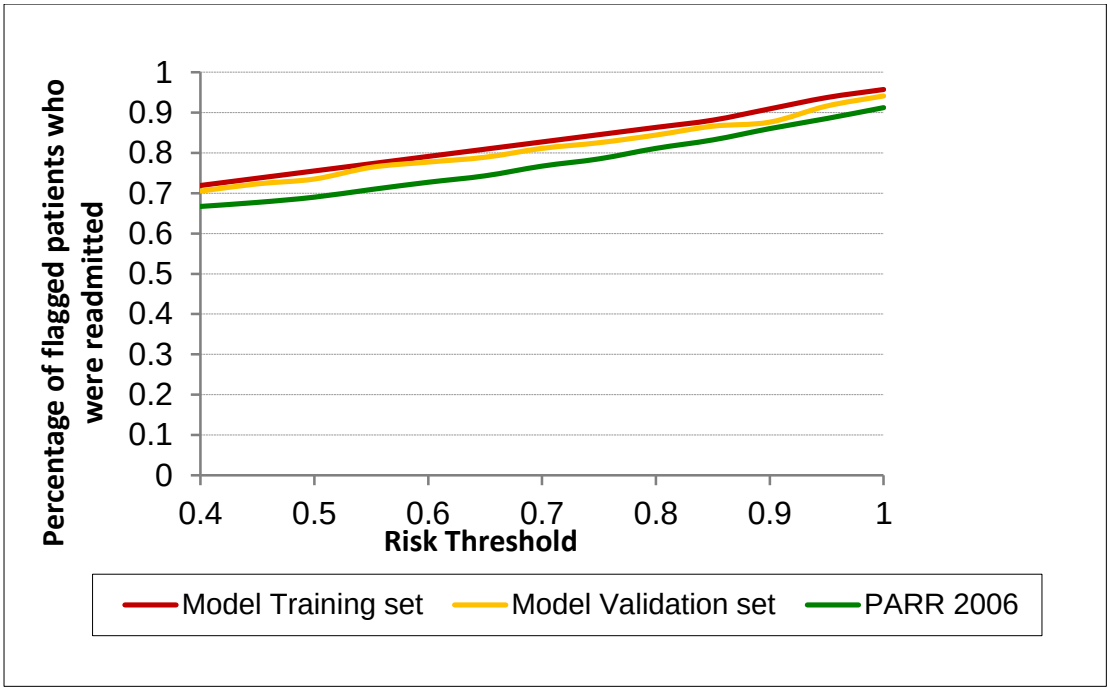

Figure A3.1 Comparison pf percentage of patients flagged by Fuzzy regression models and PARR 2006

Bibliography

Abbod, M.F., Catto, J.F., Linkens, D.A., Tan, D. (2007). Application of Artificial Intelligence to the Managementof Urological Cancer,. *The Journal of Urology,* 178(1), 1150-1156.

Abbod, M.F., Linkens, D., Von Keyserlingk, D.G, Linkens , D.A.,Mahfouf, M. (2006) Survey of utilisation of fuzzy technology in medicine and healthcare. *Fuzzy Sets and Systems,* 120(2), 331-349.

Adrion, E.R., Aucott, J., Lemke, K.W., Weiner, J.P. (2015). *Health care costs, utilization and patterns of care.* PLoS One, 10(2).

Ali, J., Khan, R. & Ahmad, N. (2012). Random Forests and Decision Trees. *IJCSI International Journal of Computer Science,* 9(5), 272-278.

Alonso-Morán, E., Nuño-Solinis, R., Onder, G., Tonnara, G. (2015). Multimorbidity in risk stratification tools to predict negative outcomes in adult population. *European Journal of Internal Medicine,* 26(3), 182-189.

Appleby, J., Harrison, T., Hawkins, L. & Dixon, A. (2012). *The King's Fund:Payment by Results.How can payment systems help to deliver better care?.* Availableat:http://www.kingsfund.org.uk/sites/files/kf/field/field_publication_file/payment-by-results-the-kings-fund-nov-2012.pdf (Accessed: 10 Jan 2016)

Acilar, A.M., Arslan, A. (2011). Optimization of multiple input-output fuzzy membership functions using clonal selection algorithm.*Expert Systemwith Applications,* 38(3), 1374-1381.

Arulchinnappan, S. & Rajendran, G. (2011). A study on reverse osmosis permeating treatment for yarn dyeing effluent using fuzzy linear regression. *African Journal of Biotechnology,* 10(78), 17969-17972.

Ashton, C. M. & Wray, N. P., (1996). A conceptual framework for the study of early readmission as an indicator of quality of care. 43(11), 1533–1541.

Auble, T. E., Hsieh, M., Gardner, W. & Cooper, G. F., (2005). A prediction rule to identify Low-risk patients with heart failure. *Academic Emergency Medicine,* 12(6), 514-521.

Au, G., Finlay, A.M., Jeffrey, A.B., Justin, E., Kaul, P. (2012). Predicting the risk of unplanned readmission or death within 30 days of discharge after a heart failure hospitalization. *American Heart Journal,* 164(3).

Austin, P. C.(2007). A comparison of regression trees, logistic regression, generalizedadditive models, and multivariate adaptive regression splines for predicting AMI mortality, *Statistics in medicine,* 26(15), 2937-2957.

Aylin P, Bottle A, Jen M.H. (2010) HSMR mortality indicators. Imperial College Technical Document. https://www1.imperial.ac.uk/resources/3321CA24-A5BC-4A91-9CC9-12C74AA72FDC/ (Accessed June 2013)

Beliakov, G.(1996). Fuzzy sets and membership functions based on probabilities. *Information Sciences,* 91(1), (95-111).

Bellman,R.E. & Zadeh, L.A. (1970). Decision making in fuzzy environment.*ManagementScience,* 17(4), 141-164

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

Billings, J., Dixon, J., Mijanovich, T. and Wennberg, D. (2006). Case finding for patients at risk of readmission to hospital: development of algorithm to identify high risk patients. *Bmj.* 333 (7563), 327.

Billings, J., Georghiou, T., Blunt, I. and Bardsley, M. (2013). Choosing a model to predict hospital admission: an observational study of new variants of predictive models for case finding. *BMJ Open.* 3 (8)

Billings, J., Mijanovich, T., Dixon, J., Curry, N., Wennberg, D., Darin, B. and Steinort, K. (2006). *Case finding algorithms for Patients at Risk of RE-Hospitalisation PARR1 and PARR2,* s.l.: NYU Center for Health and Public Service Research.

Bisserier, A., Boukezzoula, R. and Galichet, S. (2010). A revisited approach to linear fuzzy regression using trapezoidal fuzzy intervals. *Information Sciences; Inf.Sci.* 180 (19), 3653-3673.

Blunt, I., Bardsley, M., Grove, A. & Aileen, C. (2014). Classifying emergency 30-day readmissions in England using routine hospital data 2004–2010: what is the scope for reduction? *Emerg Med J,* 1-7.

Bosc, P., Lietard, L. and Pivert, O. 1995. *Fuzziness in Database Management Systems.* 5, Physica-Verlag HD, 275-308.

Bottle, A., Aylin, P. and Majeed, A. (2006). Identifying patients at high risk of emergency hospital admissions: a logistic regression analysis. *Journal of the Royal Society of Medicine.* 99 (8), 406.

Bottle, A., Gaudoin, R., Jones, S. & Aylin, P. (2014). Can valid and practical risk-prediction or casemix adjustment models, including adjustment for cormbidity, be generated from English hospital administrative data (Hospital EpisodeStatistics)? A national observational study. *Health Services and Delivery Research,* 2(40), 2050-4349.

Bradley, E.H., Curry, L., Horwitz, L.I., Sipsma, H., Thompson, J.W., Elma, M., Walsh, M.N. and Krumholz, H.M. (2012). Contemporary Evidence About Hospital Strategies for Reducing 30-Day Readmissions, A National Study. *Journal of the American College of Cardiology,* 60(7), 607-14.

Bradley, E., Yakusheva, O., Horwitz, L.I., Sipsma, H. and Fletcher, J. (2013). Identifying Patients at Increased Risk for Unplanned Readmission. *Med Care,* 51(9), 761-766.

Briefing NHS Confederation, (2011). *Foundation Trust Network, The impact of non-payment for acute readmissions.* Available at: http://www.chks.co.uk/ userfiles/ files /The%20impact%20of%20nonpayment%20for%20acute%20readmissions%20FINAL %20FOR%20WEB.pdf (Accessed: 10 March 2016)

Brunetto, A. T, Ang J.E., Olmos, D., Tan, D., Barriuso, J., Arkenau, H.T., Yap, T.A., Molife, Rhoda, L., Banerji, U., Bono, D., Judson, I. and Kaye, S. (2010). A study of the pattern of hospital admissions in a specialist Phase I oncology trials unit: Unplanned admissions as an early indicator of patient attrition. *European Journal of Cancer.* 46 (15), 2739-2745.

Byrne, D.G., Chung, S.L., Bennett, K. and Silke, B. (2010). Age and outcome in acute emergency medical admissions. *Age and Ageing.* 39 (6), 694-698.

Chan, F.W., Wong, F.Y., Yam, C.H., Cheung, W.L., Wong, E.L., Leung, M.C., Goggins, W.B. and Yeoh, E.K. (2011). Risk factors of hospitalization and readmission of patients with COPD in Hong Kong population: Analysis of hospital admission records. *BMC Health Services Research.* 11,186-186.

Charlson, M.E, Pompei, P., Ales, K.L, et al. (1987) A new method of classifying prognostic comorbidity in longitudinal studies: development and validation. *J Chronic Dis*. 40(5),373–383.

Chen, L.H. and Hsueh, C.C. (2009). Fuzzy Regression Models Using the Least-Squares Method Based on the Concept of Distance. *IEEE Transactions on Fuzzy Systems, 17(6)*, 1259-1272.

Chen, S.J. and Chen, S.M. (2003). Fuzzy Risk Analysis Based on Similarity Measures of Generalized Fuzzy Numbers. *IEEE Transacions on Fuzzy Systems,* 11(1), 45-56.

Chen, T. & Wang, Y. (2012). Long-term load forecasting by a collaborative fuzzy-neural approach. *International Journal of Electrical Power and Energy Systems,*43(1), 454-464.

Chen, T. Y. & Lai, H. L. (2011). A risk management method for enhancing patient safety based on interval-valued fuzzy numbers. *African Journal of Business Management,* 5(30), 11925-11945.

Concato, J., Peduzzi, P., Kamina, A. and Horwitz, R.I. (2001). A nested case–control study of the effectiveness of screening for prostate cancer: research design. *Journal of Clinical Epidemiology.* 54 (6), 558-564.

Coppi, R. (2008). Management of uncertainty in Statistical Reasoning: The case of Regression Analysis. *International Journal of Approximate Reasoning.* 47 (3), 284-305.

Corrigan, J. and Martin, J. (1992). Identification of factors associated with hospital readmission and development of a predictive model. *Health Services Research.* 27 (1), 81-102.

Curry, N., Billings, J., Darin, B., Dixon, J., Williams, M. and Wennberg, D. (2005). *Predictive risk project,Literature Review,* NHS Modernisation Agency.

de Groot, V., Beckerman, H., Lankhorst, G.J, Bouter, L.M. (2003) How to measure comorbidity: a critical review of available methods. *J Clin Epidemiol.* 56(3),221–229.

D'Urso, P., De Giovanni, L. and Spagnoletti, P. (2013). A fuzzy taxonomy for e-health projects.*International Journal of Machine Learning and Cybernetics,* Volume 4, 487-504.

Del Sindaco, D., Pulignano, G., Minardi, G., Apostoli, A., Guerrieri, L., Rotoloni, M., Petri, G., Fabrizi, L., Caroselli, A., Venusti, R. and Chiantera, A. (2007). Two-year outcome of a prospective, controlled study of a disease management programme for elderly patients with heart failure. *Journal of Cardiovascular Medicine (Hagerstown, Md.).* 8 (5), 324.

Demir, E., Chaussalet, T.J., Xie, H. and Millard, P.H. (2008). Emergency Readmission Criterion: A Technique for Determining the Emergency Readmission Time Window. *IEEE Transactions on Information Technology in Biomedicine.* 12 (5), 644-649.

Demir, E. (2014). A Decision Support Tool for Predicting Patients at Risk of Readmission: A Comparison of Classification Trees, Logistic Regression, Generalized Additive Models, and Multivariate Adaptive Regression Splines. *Decision Sciences.* 45 (5), 849-880.

Department of Health, (2012). *A simple guide to payment by results.* Available at: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/213150/PbR-Simple-Guide-FINAL.pdf(Accessed: 12 March 2016)

Details of episodes to spell conversation, (http://www.hscic.gov.uk/SHMI) (Accessed 10 December 2015)

Diamond, P., (1988). Fuzzy Least Squares. *Information Sciences,* Volume 46, 141-157

DiRusso, S.M., Chahine, A.A., Sullivan, T., Risucci, D., Nealon, P., Cuff, S., Savino, J. and Slim, M. (2002). Development of a model for prediction of survival in pediatric trauma patients: Comparison of artificial neural networks and logistic regression. *Journal of Pediatric Surgery.* 37 (7), 1098-1104.

Dom, R.M., Kareem, S.A., Rasak, I.A. and Abidin, B. (2008). A Learning System Prediction Method Using Fuzzy Regression. *Lecture Notes in Engineering and Computer Science.* 2168 (1), 69-72.

Dorman, R.B., Miller, C.J., Leslie, D.B., Serrot, F.J., Slusarek, B., Buchwald, H., Connett, J.E. and Ikramuddin, S. (2012). Risk for Hospital Readmission following Bariatric Surgery. *PLoS One.* 7 (3).

Foot, C., Sonola, L., Bennett, L., Fitzsimons, B., Raleigh, V. and Gregory, S. (2014). *Managing quality in community health care services.* Available at:http://www.kingsfund.org.uk/sites/files/kf/field/field_publication_file/managing-quality-in-community-health-care-services.pdf(Accessed: 14 Feb 2016)

Friedmann, P.D., Jin, L., Karrison, T.G., Hayley, D.C., Mulliken, R., Walter, J. and Chin, M.H. (2001). Early Revisit, Hospitalization, or Death among older persons discharged from the ED. *American Journal of Emergency Medicine.* 19 (2), 125-129.

Fialho, A.S., Cismondi, F., Vieira, S.M., Reti, S.R., Sousa, J.M. and Finkelstein, S.N. (2012). Data mining using Clinical physiology at discharge to predict ICU re-admissions. *Expert Systems with Applications,* Volume 39,13158–13165.

García-Pérez, L., Linertová, R., Lorenzo-Riera, A., Vázquez-Díaz, J.R., Duque-González, B. and Sarría-Santamera, A. (2011). Risk factors for hospital readmissions in elderly patients: a systematic review. *QJM: An International Journal of Medicine.* 104 (8), 639-651.

Gerds, T.A., Cai, T. and Schumacher, M.(2008). The Performance of Risk Prediction Models. *Biometrical Journal,* 50(4), 457–479.

Georghiou, T., Billings , J. & Bardsley, M. (2013) *New predictive case finding models for the NHS.*, Nuffield Trust Predictive Risk Conference 2013.

Gorzalczany, M.B. and Piasta, Z. (1999). Neuro-fuzzy approach versus rough-set inspired methodology for intelligent decision support. *An International Journal of Information Sciences,* 120(1), 45-68.

Lewis, G. (2015). *Next Steps for Risk Stratification in the NHS.* Available at: https://www.england.nhs.uk/wp-content/uploads/2015/01/nxt-steps-risk-stratglewis.pdf(Accessed: 10 Feb 2016)

Graham, L.E., Leff, B. and Arbaje, A.I. (2013). Risk of Hospital Readmission for Older Adults Discharged on Friday. *Journal of the American Geriatrics Society.* 61 (2), 300-301.

Gruneir, A., Dhalla, I. A., van Walraven, C., Fischer, H. D., Camacho, X., Rochon, P. A., & Anderson, G. M.. (2011). Unplanned readmissions after hospital discharge among patients identified as being at high risk for readmission using a validated predictive algorithm. *Open Medicine.* 5 (2), e104-e111.

Hasan, O., Meltzer, D.O., Shaykevich, S.A., Bell, C.M., Kaboli, P.J., Auerbach, A.D., Wetterneck, T.B., Arora, V.M., Zhang, J. and Schnipper, J.L. (2010). Hospital

Readmission in General Medicine Patients: A Prediction Model. *Journal of General Internal Medicine.* 25 (3), 211-219.

Hensher, M., Edwards, N. and Stokes, R. (1999). International trends in the provision and utilisation of hospital care. *Bmj.* 319 (7213), 845.

HES data dictionary from http://www.hesonline.nhs.uk/Ease/servlet/ContentServer?siteID=1937&categoryID=571 (Accessed 10 Aug 2013)

Hojati, M., Bector, C.R. and Smimou, K. (2005). A simple method for computation of fuzzy linear regression. *European Journal of Operational Research.* 166 (1), 172-184.

Hong, D. H. & Hwang, C. H., (2004). Extended fuzzy regression models using regularization method. *Information Sciences,* 164(1), 31-46.

Howell, S., Coory, M., Martin, J. and Duckett, S. (2009). Using routine inpatient data to identify patients at risk of hospital readmission. *BMC Health Services Research.* 9 96-96.

HRG4 Grouper Reference Manual, http://www.hscic.gov.uk/media/11481/HRG4-Grouper-Reference-Manual-HRG4_Grouper_Reference_Manual_v3.0.pdf (Accessed: 19 Jan 2016)

Hung, W. and Yang, M. (2006). An omission approach for detecting outliers in fuzzy regression models. *Fuzzy Sets and Systems.* 157 (23), 3109-3122.

Johns Hopkins ACG, (2014). *Applications of the ACG System in the UK.* Available at: https://www.cscsu.nhs.uk/wp-content/uploads/2015/01/ACG-White-Paper-UK-Applications.pdf(Accessed: 17 Feb 2016)

Kansagara, D., Englander, H., Salanitro, A., Kagen, D., Theobald, C., Freeman, M. and Kripalani, S. (2011). *Risk Prediction Models for Hospital Readmission A Systematic Review.*

Kasper, E.K., Gerstenblith, G., Hefter, G., Van Anden, E., Brinker, J.A., Thiemann, D.R., Terrin, M., Forman, S. and Gottlieb, S.H. (2002). A randomized trial of the efficacy of multidisciplinary care in heart failure outpatients at high risk of hospital readmission. *Journal of the American College of Cardiology.* 39 (3), 471-480.

Kayacan, E., Ulutas, B. & Kaynak, O. (2010). Grey system theory-based models in time series prediction. *Expert Systems with Applications,* Volume 37, 1784-1789.

Kim, K. J., Moskowitz, H., Koksalan, M. (1996). Fuzzy versus statistical linear regression. *European Journal of Operational Research,* 92,417-434.

King's Fund, (2006). *Patients at risk of re-hospitalisation (PARR) case finding tool.* Available at: http://www.kingsfund.org.uk current projects/predictive risk/index.html. (Accessed: 10 June 2014)

Krumholz, H.M., Parent, E.M., Tu, N., Vaccarino, V., Wang, Y., Radford, M.J. and Hennen, J. (1997). Re-admission after hospitalizationfor congestive heart failure among Medicare beneficiaries. *Arch Int MED,* 157, 99-104.

Kunjunninair, A. P. (2012). Clinical decision support system: risk level prediction of heart disease using weighted fuzzy rules and decision tree rules. *Central European Journal of Computer Science,* 2(1), 86-86.

Lagoe, R.J., Nanno, D.S. and Luziani, M.E. (2012). Quantitative tools for addressing hospital readmissions. *BMC Research Notes.* 5 (1).

Lemon, S.C., Roy, J., Clark, M.A., Friedmann, P.D. and Rakowski, W. (2003). Classification and regression tree analysis in public health: Methodological review and comparision with logistic regression. *Annals of Behavioral Medicine,*26, 172-181.

Lewis, G. (2015). *Next Steps for Risk Stratification in the NHS.* Available at: https://www.england.nhs.uk/wp-content/uploads/2015/01/nxt-steps-risk-strat-glewis.pdf (Accessed: 16 March 2016)

Lin, C.C., Ou, Y.K., Chen, S.H., Liu, Y.C. and Lin, J. (2010). Comparison of artificial neural network and logistic regression models for predicting mortality in elderly patients with hip fracture. *Injury.* 41 (8), 869-873.

Li, T., Jin, J. and Li, C. (2012). Refractured Well Selection for Multicriteria Group Decision Making by Integrating Fuzzy AHP with Fuzzy Topsis Based on Interval-Typed Fuzzy Numbers. *Journal of Applied Mathematics,* 1-21.

Liu, P., Lei, L., Yin, J., Zhang, W., Naijun, W. and El-Darzi, E. (2006). *Healthcare Data Mining: Prediction Inpatient Length of Stay.*

Lohani, A.K., Goel, N.K. and Bhatia, K.K.S. (2006). Takagi–Sugeno fuzzy inference system for modeling stage–discharge relationship. *Journal of Hydrology,* 331(1), 146-160.

Lundin, M., Lundin, J., Burke, H.B., Toikkanen, S., Pylkkänen, L. and Joensuu, H. (1999). Artificial neural networks applied to survival prediction in breast cancer. *Oncology.* 57 (4), 281.

Marcantonio, E.R., McKean, S., Goldfinger, M., Kleefield, S., Yurkofsky, M. and Brennan, T.A. (1999). Factors associated with unplanned hospital readmission among patients 65 years of age and older in a medicare managed care plan. *The American Journal of Medicine.* 107 (1), 13-17.

McCauley-Bell, P.R., Crumpton, L.L. and Wang, H. (1999). Measurement of cumulative trauma disorder risk in clerical tasks using fuzzy linear regression. *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions On.* 29 (1), 1-14.

Meenan, R.T., Goodman, M.J., Fishman, P.A., Hornbrook, M.C., O'Keeffe-Rosetti, M.C. and Bachman, D.J. (2003). Using Risk-Adjustment Models to Identify High-Cost Risks. *Medical Care.* 41 (11), 1301-1312. Möller, B., Graf, W., Beer, M. & Sickert, J., 2002. Fuzzy Randomness - Towards a new Modeling of Uncertainty. *WCCM V Fifth World Congress on Computational Mechanics.*

Möller, B., Graf, W., Beer, M. & Sickert, J. (2002). Fuzzy Randomness - Towards a new Modeling of Uncertainty. *WCCM V Fifth World Congress on Computational Mechanics.*

Moran, J., Colbert, C.Y., Song, J., Hull, J., Rajan, S., Varghees, S., Arroliga, A.C. and Reddy, S.P. (2013). Residents Examine Factors Associated with 30-Day, Same-Cause Hospital Readmissions on an Internal Medicine Service. *American Journal of Medical Quality.* 28 (6), 492-501.

Motro, A. (1995). Imprecision and uncertainty in Database Systems. *Fuzziness in Database Management Systems,* 3-22.

Muzzarelli, S., Leibundgut, G., Maeder, M.T., Rickli, H., Handschin, R., Gutmann, M., Jeker, U., Buser, P., Pfisterer, M., Brunner-La Rocca, H.P. and TIME-CHF Investigators. (2010). Predictors of early readmission or death in elderly patients with heart failure. *American Heart Journal.* 160 (2), 308-314.

Nagar, P. & Srivastava, S. (2008). Adaptive Fuzzy Regression Model for the Prediction of Dichotomous Response Variables using Cancer Data: A Case Study,. *Journal of Applied Mathematics and Informatics (JAMSI),* 4(2), 183-191.

NHS England, (2015). *Case Finding & Risk Stratification Handbook.* Available at: https://www.england.nhs.uk/wp-content/uploads/2015/01/2015-01-20-CFRS-v0.14-FINAL.pdf (Accessed: 16 March 2016)

Nuffield Trust, (2011). *Predictive risk and health care: an overview. London: Nuffield Trust.*

Ottenbacher, K.J., Linn, R.T., Smith, P.M., Illig, S.B., Mancuso, M. and Granger, C.V. (2004). Comparison of logistic regression and neural network analysis applied to predicting living setting after hip fracture. *Annals of Epidemiology.* 14 (8), 551-559.

Parker, J.P., McCombs, J.S. and Graddy, E.A. (2003). Can Pharmacy Data Improve Prediction of Hospital Outcomes? Comparisons with a Diagnosis-Based Comorbidity Measure. *Medical Care.* 41 (3), 407-419.

Patel, P. & Marwala, T. (2011). Fuzzy inference systems optimization.

Pavon, J.M., Zhao, Y., McConnell, E. and Hastings, S.N. (2014). Identifying Risk of Readmission in Hospitalized Elderly Adults Through Inpatient Medication Exposure.

Pearson, B., Skelly, R., Wileman, D. and Masud, T. (2002). Unplanned readmission to hospital: a comparison of the views of general practitioners and hospital staff. *Age and Ageing.* 31 (2), 141.

Peters, G., (1994). Fuzy Linear Regression with fuzzy intervals. *Fuzzy Sets and Systems,* 63, 45-55.

Philbin, E.F. and Disalvo, T.G. (1999). Prediction of hospital readmission for heart failure: development of a simple risk score based on administrative data. *Journal of the American College of Cardiology.* 33 (6), 1560-1566.

Phuong, N.H. and Kreinovich, V. (2001). Fuzzy logic and its applications in medicine. *International Journal of Medical Informatics,* 62(2), 165-173.

Ponzo, M. & Scoppa, Vincenzo. (2016). Cost-Sharing and Use of Health Services in Italy: Evidence from a Fuzzy Regression Discontinuity Design. *IZA working paper,* Issue 9772

Pourahmad, S., Ayatollahi, S.M.T., Taheri, S.M. and Agahi, Z.H. (2011). Fuzzy logistic regression based on the least squares approach with application in clinical studies. *Computers and Mathematics with Applications.* 62 (9), 3353-3365.

Provotar, A.I., Lapko, A.V. and Provotar, A.A. Fuzzy Inference Systems and their Applications. *Cybernetics and Systems Analysis,* 49(4), 517-525.

Purdy, S. (2010). *Avoiding hospital admissions. What does the research evidence say?* Available at: http://www.kingsfund.org.uk/sites/files/kf/Avoiding-Hospital-Admissions-Sarah-Purdy-December2010.pdf (Accessed: 12 Jan 2016)

Ramesh, A.N., Kambhampati, C., Monson, J.R.T. and Drew, P.J., (2004). Artificial intelligence in medicine. *Annals of The Royal College of Surgeons of England*, *86*(5), 334-8.

Ramli, A.A., Watada, J. and Pedrycz, W. (2011). Real-time fuzzy regression analysis: A convex hull approach.. *European Journal of Operational Research,* 210(3), pp. 606-617.

Rasmusson, K., Bunizillo, J., Kfoury, A.G., Lappe, D., Alharethi, R., Horne, B., Nixon, J. and Budge, D. (2013). *A Predictive Model of Heart Failure Readmissions: Results of a Multivariate Analysis.*

Razi, M.A. and Athappilly, K. (2005). A comparative predictive analysis of neural networks (NNs), nonlinear regression and classification and regression tree (CART) models. *Expert Systems with Applications.* 29 (1), 65-74.

Riegel, B. et al. (2002). Effect of a standardized nurse case-management telephone intervention on resource use in patients with chronic heart failure. *Archives of Internal Medicine.* 162 (6), 705.

Sener, Z. & Karsak, E. E., (2011). A combined fuzzy linear regression and fuzzy multiple objective programming approach for setting target levels in quality function deployment. *Expert Systems with Applications,* 38(4), pp. 3015-3022.

Shakouri G, H. & Nadmi, R. (2009). A novel fuzzy linear regression model based on a non-equality possibility index and optimum uncertainty. *Journal of Applied Soft Computing,* 9(2), pp. 590-598.

Setiawan, N. A. (2014). Fuzzy Decision Support System for Coronary Artery Disease Diagnosis Based on Rough Set Theory. *International Journal of Rough sets and Data analysis, 1(1), 65-80.*

Sg2, (2011). Sg2 Service Kit: Reducing 30-Day Emergency Readmissions. Available at:*http://www.hsj.co.uk/Journals/2/Files/2011/6/15/Sg2_Service%20Kit_Reducing%2030-Day%20Readmissions.pdf*(Accessed: 12 Jan 2016)

Shapiro, A. F. (2005). Fuzzy Regression models. *ARC.* Available at: *http://alm.soa.org/library/research/actuarial-research-clearinghouse/2006/january/arch06v40n1-ii.pdf* (Accessed: 15 Jan 2013)

Sharon, E., Krebs, C., Turner, W., Desai, N., Binus, G., Penk, W. and Gastfriend, D.R. (2004). Predictive Validity of the ASAM Patient Placement Criteria for Hospital Utilization. *Journal of Addictive Diseases,* 22, 79-93.

Solomatine, D.P. and Shrestha, D.L. (2009). A novel method to estimate model uncertainty using machine learning techniques. *Water Resources Res,* Volume 45.

Steimann, F. (2001). On the use and usefulness of fuzzy sets in medical AI. *Artificial Intelligence in Medicine,* 21(1),131-137.

Su, C.T., Yang, C.H., Hsu, K.H. and Chiu, W.K. (2006). Data mining for the diagnosis of type II diabetes from three-dimensional body surface anthropometrical scanning data. *Computers & Mathematics with Applications,* 51(6-7), 1075-1092

Su, Z.G., Wang, P.H. and Song, Z.L. (2013). Kernel based nonlinear fuzzy regression model. *Engineering Applications of Artificial Intelligence.* 26 (2), 724-738.

Taheri, S.M. and Kelkinnama, M. (2008). *Fuzzy least absolutes regression.* Varna,Bulgaria, IEE Takemura, K., 2004. Fuzzy logistic regression analysis for fuzzy input-output data. *The Proceedings of the Second International Symposium on Soft Computing and Intelligent System,* pp. 1-6.

Takeuti, G. and Titani, S. (1984). Intutionistic Fuzzy Logic and Intuitionistic Fuzzy set Theory.*The Journal of Symbolic Logic,* 49(3), 851.

Tamaki, F., Kanagawa, A. and Ohta, H. (1998). Identification of membership functions based on fuzzy observation data. *Fuzzy Sets and Systems,* 93(3), 311-318.

Tanaka , H., (1987). Fuzzy Data Analysis by Possibilistic Linear Models. *Fuzzy Sets and Systems,* Volume 363-375, 24.

Tanaka, H., (1989). Possibilistic linear regression analysis for fuzzy data. *European Journal of Operational Research,* 40(3),389-396.

Tanaka, H. and Watada, J. (1989). Possibilistic Linear Systems and their application to Linear Regression Models. *Fuzzy Sets and Systems,* Volume 27, 275-289.

Tsien, C.L., Fraser, H.S., Long, W.J. and Kennedy, R.L. (1998). Using classification tree and logistic regression methods to diagnose myocardial infection. *Medinfo,* Volume 9, 493-497.

Takehira, R., Murakami, K., Katayama, S., Nishizawa, K. and Yamamura, S. (2011). Artificial Neural Network Modeling of Quality of Life of Cancer Patients: Relationships between Quality of Life Assessments, as Evaluated by Patients, Pharmacists, and Nurses. *International Journal of Biomedical Science: IJBS.* 7 (4), 255-262.

Wang, T.N., Cheng, C.H. and Chiu, H.W. (2013). *Predicting post-treatment survivability of patients with breast cancer using Artificial Neural Network methods.*

Thomson, K. & Lewis, G. (2013). *Information Governance and Risk Stratification: Advice and Options for CCGs and GPs.* Available at: http://ivygrove.org.uk/downloads/CD-ig-risk-ccg-gp.pdf(Accessed: 15 Jan 2016)

Tsien, C.L., Fraser, H.S., Long, W.J. and Kennedy, R.L. (1998). Using classification tre and logistic regression methods to diagnose myocardial infection. *Medinfo,* Volume 9, 493-497.

Tsipouras, M.G., Exarchos, T.P., Fotiadis, D.I., Kotsia, A.P., Vakalis, K.V., Naka, K.K. and Michalis, L.K. (2008). Automated Diagnosis of Coronary Artery Disease based on Data Mining and Fuzzy Modelling. *IEEE Transactions on Information Technology in BioMedicine,* 12(4).

Tu, J.V. (1996). Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes. *Journal of Clinical Epidemiology.* 49 (11), 1225-1231.

Van Walraven, C., Bennett, C., Jennings, A., Austin, P.C. and Forster, A.J. (2011). Proportion of hospital readmissions deemed avoidable: a systematic review. *CMAJ: Canadian Medical Association Journal = Journal De L'Association Medicale Canadienne.* 183 (7), E391.

Vilaró, J., Ramirez-Sarmiento, A., Martínez-Llorens, J.M., Mendoza, T., Alvarez, M., Sánchez-Cayado, N., Vega, Á., Gimeno, E., Coronell, C., Gea, J. and Roca, J. (2010). Global muscle dysfunction as a risk factor of readmission to hospital due to COPD exacerbations. *Respiratory Medicine.* 104 (12), 1896-1902.

Wakefield, B.J., Ward, M.M., Holman, J.E., Ray, A., Scherubel, M., Burns, T.L., Kienzle, M.G. and Rosenthal, G.E. (2008). Evaluation of home telehealth following hospitalization for heart failure: a randomized trial. *Telemedicine Journal and E-Health: The Official Journal of the American Telemedicine Association.* 14 (8), 753-761.

Wang, T. N., Cheng, C. H. & Chiu, H. W., (2013). *Predicting post-treatment survivability of patients with breast cancer using Artificial Neural Network methods.* Osaka, Engineering in Medicine and Biology Society (EMBC), 2013 35th Annual International Conference of the IEEE.

Weir, S. & Jones, W. C. (2009). *Overview and Uses of Predictive Modeling,* Center for Health Policy and Research.

Wei, S. H. & Chen, S. M. (2009). Fuzzy risk analysis based on interval-valued fuzzynumber. *Expert Systems with Applications,* 36(2), 2285-2299.

Yager, R. R. (1986). On the Theory of Bags. *International Journal of General Systems,* Volume 13, 23-37.

Yang, M. S. & Liu, H. H. (2003). Fuzzy least-squares algorithms for interactive fuzzy linear regression models. *Fuzzy Sets and Systems,* Volume 135, 305-316.

Yang, X., Peng, B., Chen, R., Zhang, Q., Zhu, D., Zhang, Q.J., Xue, F. and Qi, L. (2014). Statistical profiling methods with hierarchical logistic regression for healthcare providers with binary outcomes. *Journal of Applied Statistics,* 41(1), 46-59.

Zadeh, L. A. (1965). Fuzzy Sets. *Information and Control,* 338-353.

Zadeh, L. A. (2005). Toward a generalized theory of uncertainty (GTU)—an outline. *Information Sciences,* 172(1), 1-40.

Zernikow, B., Holtmannspötter, K., Michel, E., Hornschuh, F., Groote, K. and Hennecke, K.H. (1999). Predicting length-of-stay in preterm neonates. *European Journal of Pediatrics; Eur.J. Pediatr.* 158 (1), 59-62.

Zhao, Y., Ash, A.S., Haughton, J. and McMillan, B. (2003). Identifying Future High-Cost Cases Through Predictive Modeling. *Disease Management & Health Outcomes.* 11 (6), 389-397.