Applied statistical methods for prediction modelling of upper limb functional recovery after stroke

Ahmad Najim Sheet Al-Shallawi

This research is submitted in partial fulfilment of the requirements for the degree of

Doctor of Philosophy

July 2019

Keele University

Contents

1 LIST OF FIGURES

2 LIST OF TABLES

3 List of Abbreviations

Abbreviation	Description
ADL	Activity of Daily Living
ARAT	Action Research Arm Test
LASSO	Least Absolute Shrinkage and Selection Operator
ALASSO	Adaptive of Least Absolute Shrinkage and Selection Operator
AIC	Akaike Information Criteria
BBT	Box and Block Test
BI	Barthel Index
BIC	Bayesian Information Criteria
CV	Cross- Validation
Cp	Mallows Criteria
CNS	Central Nervous System
CIMT	Central Induced Movement Daly's Disability-Adjusted Life Years
FP	False Positive
FN	False Negative
GLASSO	Group Absolute Square shrinkage of operator
ICF	International Classification Functions
LACS	Lacunar Syndromes
MI	Motricity Index
MRS	Modified Rankin Scale
NIHSS	National Institute of Stroke Scale
NB	Net Benefit
NIS	National Health Service
OCSP	Oxfordshire Community Stroke
ROC	Receiver of Operators' Curve
POCS	Posterior Circulation Syndromes
PACS	Partial Anterior Circulation Syndromes
TP	True Positive

TN	True Negative
TACS	Total Anterior Circulation Syndromes
3m	Model after three months
6m	Model after three months
WHO	World Health Organisation

4 ACKNOWLEDGEMENTS

I am very grateful to my supervisor, Professor Anand Pandyan, and co-supervisor, Dr. Dimitra Blana, for the knowledge and skills they have equipped me with. I am thankful for our research team meetings (Dr. Ed Chadwick, Caroline Stewart and Charles Day) as this work could not be a reality without their encouragement and expert guidance. It had been a pleasure to share my student experience at the MacKay building, with my faithful colleagues Ali, Amnah, Aseel, Fraser, Mohmmad, Hamzah, and Shallum. The intellectual and inspiring conversation will never be forgotten.

I extend my gratitude and appreciation to my family, particularly my mother and sisters, who I have missed since departing from them at the beginning of my PhD journey. I would like to thank my wife, Huda, for her support, delicious meals, continuing love and support throughout this period of my project and of course, to my son, Fahd, and my daughters, Rahma and Salma, for their understanding, support and patience; they always make me happy and put a smile on my face. The work in this research was funded by the Ministry of Higher Education and Scientific Research in Iraq which I gratefully acknowledge.

5 Abstract

Stroke is the third largest cause of death in the world, with a significant contribution to disability. Motor function impairment, encompassing upper limb impairment, is the most significant post-stroke impairment. Such an impairment contributes to reducing a person's ability to complete daily activities, thus affecting their quality of life. Effective interventions, specifically targeted at upper limb recovery, are important, just as much as predictions of patient's post-stroke. Predictions have become essential in making accurate clinical decisions in stroke management, including selection of appropriate rehabilitation programs, referring into appropriate services, setting realistic goals by therapists and clinicians and predicting the level of dependence following discharge from the hospital. This research focuses on the prediction of upper limb recovery and function. Despite the current and widely used traditional statistical methods of prediction, the research here presents a developed modern method which focuses on prediction models of regression methods. This is because traditional methods have been shown to lack clinical usefulness and do not have meaningful acceptance in clinical practice. The modern method developed and adopted aims to give more beneficial and valid results from the prediction model.

Chapter One

1 Introduction

1.1 Introduction to the study and its context

This study examines important elements of how statistics and medicine come together to form accurate and reliable predictions of patients with stroke. This research examines a section of concern within post-stroke patients: upper limb functional recovery. This is directly linked to carrying out activities of daily living (ADL). It examines rehabilitation and recovery, considered within the International Classification Functions Framework (ICFF) for measurement. It provides a review of the different measures of upper limb motor function, including Action Research Arm Test (ARAT), Fugl-Meyer, Wolf motor function test, Box and Blocks test (BBT). It focuses mainly on ARAT as the outcome measure (dependent variable). This research successfully sets a new cut off point for ARAT, and not only uses the modern method of Least Absolute Shrinkage and Selection Operator (LASSO) in developing a prediction model for upper limb recovery post, but it has also tested its external validation. Additionally, the results of the new model were compared with a traditional method (stepwise method). The adaptive LASSO (ALASSO) was found to be the best method with respect to performance. Decision-analytic measure was used to summarise the performance of the model in support of decision making. It is worth mentioning that this is the first time a decision-analytic measures method has been used in stroke related studies, and so this is a novelty of this thesis, contributing to knowledge in biostatistics.

1.2 Use of terminology in this research

The abbreviations for some key terminologies within biostatistics are mentioned previously and these will be referred to throughout my research. There are some key terms that I will be using in this research, that are used interchangeably such as classical and traditional methods, binary or dichotomous etc. For simplicity, the terms classical, traditional or subset selection methods will be interchangeably used to describe the traditional methods of predictors selection; whereas, penalised, modern or regularised selection methods will be interchangeably used to describe the modern methods of predictors selection such as LASSO.

1.3 Rationale for the study

As a professional statistical programmer and analyst, the author has always been fascinated by the ways in which mathematical and analytical software can be used within the medical field to predict, prevent and improve the life of existing patients. This has affected me on a personal level, after losing my son at a young age due to brain damage. As a researcher, I am constantly looking at ways to improve prediction and applied in biostatistics in clinical settings for the greater development of medicine and this has inspired me to take on this research project. Considering the insufficient validity of current predictive models of upper limb functional recovery after stroke used in clinical decision-making setting and the impracticality of using the current models in a clinical setting, I believe it is necessary to establish additional predictive models that, when coupled with clinical assessment, can improve prediction precision. As suggested from development and validation studies(Kwah and Herbert, 2016), the only current model for arm recovery (the proportional recovery model) that has been externally validated, does not give a good prediction of recovery for all patients with stroke. This model is limited and appears to predict outcomes in people with less severe strokes(Kwah and Herbert, 2016).

Therefore, providing an accurate and robust model which is well developed and externally validated, prior to its use in clinical practice, would provide a more efficient and sensitive method of prediction. This would help identify patients who are more likely to recover and assist in directing available resources toward achieving treatment goals, again contributing to the better development within medicine.

1.4 Research aims and objectives

The main aim of this research is:

- **To develop and improve a prediction model of recovery for upper limb function post-stroke.**

This aim will be achieved through four primary objectives:

1) Modify a cut-off point for the action research arm test, selected as the outcome/ dependent variable, using cluster analysis as an assistive tool in developing a prediction model.
2) Test and identify predictor variables which have a strong relationship with the dependent variable, using classical and modern methods of selection.
3) Test external validation of the models and present the benefit of each type of model that is developed based on traditional and modern methods.
4) Develop a model to determine an essential effect predictor in intervention model.

1.5 Research questions

My research aims at answering the following three research questions:

1. Should the cut-off point of the Action Research Arm Test (ARAT) outcome(s) be modified?

2. What are the primary predictors of patients' recovery post-stroke, and why?

3. Can the current methods be improved and developed using a new method of modelling?

1.6 Chapter summary

This chapter introduces the area of the research. It also sets the scene for the research, on various levels, from the basis of my personal interest and stance as researcher. It gives information on the aims and objectives of the research as well as the research questions.

Chapter two provides a review of the literature with respect to stroke as a medical condition, as well as statistical tools involved in the development of models and cluster analysis. It provides some insights into and critical points of prediction models of upper limb/arm post-stroke. This will allow the determination of the most common predictors that are used in previous prediction studies. It will also help determine the effectiveness of statistical models used in previous studies and limitations of the previous prediction models and decisions.

Chapter three begins by reviewing and describing the two types of statistical tools: regression analysis (with a review of the traditional and penalised methods of model selection in logistic regression models and assessment methods) and cluster analysis.

In **chapter four**, the means by which modification of the cut-off point of the dependent variable (outcome) is discussed for a logistic regression model

Chapter five involves applying traditional and penalised model selection methods and compares the performances of traditional and penalised methods based on multiple logistic regression.

The new application work in **chapter six** focuses on modifying the cut-off point using statistical investigation to compare the achievements of traditional and penalised methods in external validation stage and the decision analysis curve with net benefit.

Chapter six also provides information on how the research achieved external validation, while **chapter seven** provides details on developing a model, which is a novelty of my research. The research will conclude in **chapter eight** with a general discussion, limitations of the work, a summary and possible future work within this field.

Chapter two

2 Literature review

The literature review begins with information on stroke, followed by a detailed description of upper limb and recovery terms, functional recovery and rehabilitation of upper limb, which is explained based on the International Classification Functions (ICF) framework. It then provides the definitions, descriptions and measurements of arm recovery, followed by a search strategy that includes an electronic search to identify studies that are linked to the inclusion criteria of this study. A detailed description of the prediction model and predictors is then provided to show how this process offers a benchmark against which predictive modelling studies of arm recovery can be evaluated. I finally discuss some critical points, concluding with a summary about prediction methods of modelling studies.

2.1 Stroke as a medical condition

According to the World Health Organisation (WHO), a stroke is defined as *"rapidly developing clinical signs of focal (or global) disturbance of cerebral function, lasting more than 24 hours or leading to death, with no apparent cause other than that of vascular origin"* (Sacco et al., 2013; Veerbeek et al., 2014). Stroke incidence in the UK has been estimated to be 257.4 per 100,000 of the population for the year (2013/2014). Stroke remains a leading cause of mortality and long-term disability, killing an estimated 650,000 people annually. Mortality rate is higher in females (23,060) than males (16,224) (Stroke Association, 2018). The total direct cost of stroke in Europe was estimated to be 50 million in 2015 (Europe, 2017) .

Stroke is a common cause of death worldwide. In those who survive, a stroke can cause significant disability (Dacosta-Aguayo et al., 2014). Stroke is also the most common cause of disability and the disability-adjusted life years (DALYs) lost due to strokes is estimated to be 4.1% of global DALYs (Murray et al., 2012). In 1990, stroke was fifth in the DALYs league table and by 2010, it had reached third position (Murray et al., 2012). In the UK, approximately 33% of stroke survivors remain functionally dependent at one-year post-stroke. Residual symptoms and increased dependence following a stroke can remain throughout a stroke patient's life (Aziz et al., 2008). The direct cost of stroke to the National Health Service (NHS) is around £8,490,000 a year. This figure is very likely to grow due to the ageing population demographics of the UK (Mortimer and Green, 2015). The impact of stroke in Europe is also significant (Europe, 2017).

Stroke can be broadly classified as ischaemic or haemorrhagic in nature. Accounting for approximately 85% of reported strokes, ischaemic strokes occur immediately after a cerebral artery becomes partially or totally blocked, decreasing tissue perfusion. Tissue perfusion is the amount of blood that a tissue is receiving from the circulation (Hennerici, 2004). Decreased tissue perfusion can lead to tissue death. Haemorrhagic stroke accounts for around 15% of all strokes. Here, a rupture of a cerebral vessel leads to an intracranial haemorrhage and raised intracranial pressure. This ultimately leads to the compression of surrounding neuronal tissue and in many cases, cell death. Therefore, the impact of strokes, although varied, can be devastating. As a consequence of stroke, residual neurological deficits can include the loss and impairment of the motor or control functions of one side of the body, such as paresis; difficulty in speech (dysphasia); decreased mental functions (cognitive) and the impairment of emotional functions.

A well-known classification method is the Oxfordshire Community Stroke Project (OCSP). This is a simple clinical method, originally devised for patients with first time strokes, to subdivide acute strokes. Based on severity of the symptoms, strokes can be classified as (Mead et al., 2000):

- Lacunar syndromes (LACS): this includes pure motor stroke, pure sensory stroke, sensorimotor stroke and ataxic hemiparesis.
- Posterior circulation syndrome (POCS): this include patients with brain stem or cerebellar signs, and/or isolated homonymous hemianopia.
- Total anterior circulation syndromes, (TACS): this includes patients presenting with the triad of hemiparesis (or hemisensory loss), dysphasia (or other new higher cortical dysfunction) and homonymous hemianopia.
- Partial anterior circulation syndrome (PACS): this involves patients presenting with only two of the features of TACS, or isolated dysphasia or parietal lobe signs.

Patients are classified as “syndromes” (TACS, PACS, LACS, and POCS), unless brain imaging has excluded intracerebral haemorrhage. In the latter case, patients are reclassified as total or partial anterior circulation infarct (TACI or PACI), lacunar infarct (LACI), and posterior circulation infarct (POCI)(Amarenco et al., 2009). All these deficits have an impact on the subjects’ ability to perform activities important for daily living, as simple as eating, dressing themselves and writing.

One of the most common post-stroke deficits is motor impairment of the upper limb (Pollock et al., 2015), which is the focus of this research. The most common subtype of stroke is damage to the middle cerebral artery that supplies the upper limb. Hence, disability of the upper limb is the most common (Balaban et al., 2011; Levin et al., 2009). Most post-stroke patients (50% to 80 %) are likely to have an impairment affecting one arm. These patients are likely to use compensation strategies to remain independent. Of these patients with impairment, 66% will only partially recover and thus require ongoing care to complete their daily activities (Feys et al., 2000a). Post-stroke, the upper limb impairments are a considerable problem and have a significant impact on stroke-related disability. Additionally, upper limb impairment has been associated with a reduction in quality of life and unhappiness (Pollock et al., 2015). The recovery of upper limb movement and function is therefore a main concern for patients, as well as professionals who deliver health services and treatments for patients suffering from a stroke (Beebe and Lang, 2009). Therefore, prediction of patient recovery would be fundamental in supporting recovery of post-stroke patients.

The literature has provided an indication on how prediction of patients' recovery post-stroke would be beneficial. Firstly, it could guide the patient’s stroke management, helping in appropriate selection of a rehabilitation program, which would allow professionals such as therapists and clinicians to set realistic and directed goals (Kwah and Herbert, 2016). Secondly, it can be used as an effective device to correctly inform patients, as well as their relatives, on the patient's situation and the actual cost-effectiveness of rehabilitation required, giving a tangible value (Eghidemwivbie and Schneeweis, 2010; Kwakkel and Kollen, 2013; Woldag et al., 2006). However, despite the existence of numerous prediction models in the field of strokes, the prediction of recovery in stroke patients is still lacking related clinical usefulness and hence considered to be inaccurate(Kwah and Herbert, 2016).

In light of the insufficient validity of current predictive models of upper limb functional recovery used in clinical decision-making setting and the impracticality of using the current models in a clinical setting, it is necessary to establish additional predictive models that, when coupled with clinical assessment, can improve prediction precision. (Kwah and Herbert, 2016).

2.2 Upper limb focus

Most stroke patients suffer from impairment of motor and other functions of upper limb. This includes sensory impairment, abnormal muscle activation patterns, reduced muscle strength and reduced functional use of the upper limb. Upper limb activities, including movement range and the gross motion of the proximal shoulder, elbow and wrist joints to fine finger dexterity for manipulating of objects, are often more affected by stroke than the lower limb functions. The patient's ability to live independently and carry out daily activities relies heavily upon the extent of motor impairment, motor functional recovery and development of compensation strategy post-stroke (Feys et al., 2000a; Feys et al., 2000b; Stinear, 2010).

2.2.1 Upper limb rehabilitation

Stroke rehabilitation is generally described as being an active, dynamic and continuing process focussed on physical, social and psychological aspects of health. Stroke rehabilitation aims to reduce the consequences of stroke, enhance patients' abilities to perform daily activities and improve quality of life (participation). According to the NICE guideline on Long-Term Rehabilitation after stroke (2013), there are different types of rehabilitation program. Examples include cognitive, vision rehabilitation and motor control (approaches can be face, upper and lower limb movements). Rehabilitation can take place at the level of the impairment; the belief is that by improving the impairments one improves activity, and this can lead to improvements in participation. Rehabilitation can teach compensatory strategies, provide assistive devices and/or modify the environment to improve activity and participation Figure 2-1.

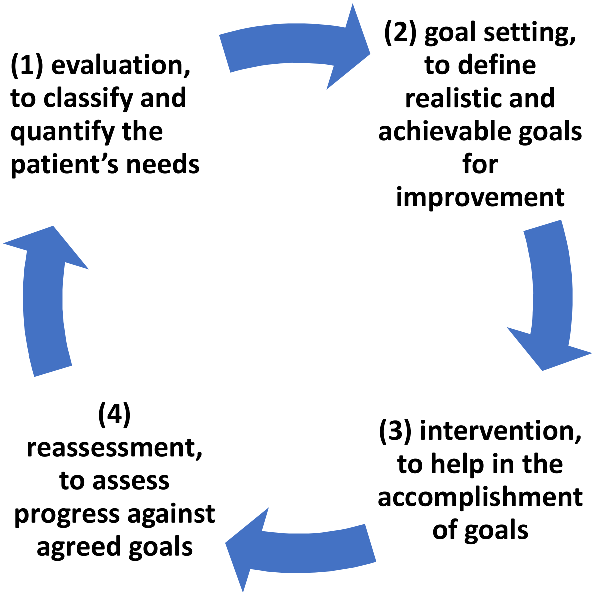

(1) evaluation, to classify and quantify the patient’s needs
(2) goal setting, to define realistic and achievable goals for improvement
(3) intervention, to help in the accomplishment of goals
(4) reassessment, to assess progress against agreed goals

Figure 2-1 Stroke rehabilitation process includes a therapeutic activity cycle.

There seems to be a moderate non-linear relation between impairment and function. More specifically, there is a scarcity in evidence to motor impairments recovery from impairment-focused therapies, which is not necessarily reflected as neurological compensation in the brain (Pinter and Brainin, 2012).

There is clear evidence at present that shows that task-based training can help functional recovery(Veerbeek et al., 2014). This corroborates with the idea that functional recovery is a result of combination of compensation and true recovery. Therefore, most rehabilitation interventions seem to work best at the level for which they are targeted. These levels are: 1) exercise treatment interventions, 2) increased amount of focused therapy or interventions compared with a reference group, 3) sensorimotor training, 4) electrical stimulation alone, biofeedback alone, or electrical stimulation in combination with biofeedback and 5) Constraint Induced Movement Therapy (CIMT) (Enderby et al., 2017).

All in all, the rehabilitation of the upper limb is a complex process which includes the retraining of gross and fine movement control of shoulder, arm, and hand. The rehabilitation program targeting hand and arm functions after stroke has lower recovery rate than that of the lower limb. Because current protocols in hospitals focus on impairments or activities that focus on mobility and transfers, the upper limb gets very little attention. Patients are often discharged before they have been fully rehabilitated (possibly due to scarcity in resources). Additionally, the complex nature of upper limb function, that requires re-learning of very fine movements patterns, is difficult to produce positive recovery results(Coupar et al., 2012). On the contrary, lower limb rehabilitation, such as gait re-education, can be achieved by re-learning gross motor skills. To improve post-stroke upper limb interventions and services that support recovery, it is vital to try developing prediction model(s) for recovery of upper limb impairments to assist in clinical testing.

2.3 Recovery

The term 'recovery' has been used to describe the processes of relearning of skills that are lost post-stroke as well as the improvement of function, regardless of how these may have occurred (Levin et al., 2009). Recovery after a stroke relies on many variables. These include:

a) the specific site of the brain damage,

b) the general health of the patient,

c) age,

d) related and unrelated diseases,

e) personality,

f) family support,

g) the care received.

Recovery after stroke has also been defined on three different levels, which are discussed later within International Classification Functions (ICF) framework (Levin et al., 2009). However, the exact mechanism and time course of upper limb post stroke injuries are not yet well investigated. Unsurprisingly, immediately after the injury, the central nervous system (CNS) falls into a period of shock (Pandyan et al., 2018). Subsequently, the CNS is believed to begin compensating for the contralesionally tilt of posture and increase loading of the ipsilesional side (Barra et al., 2009). This is then followed by a period of neuroplasticity. However, the period of neuroplasticity could not be estimated. For example, neuroplasticity may go for a long time after stroke or it might be possible to increase the opportunity for plasticity in the early stages of CNS reorganisation. This period depends on factors such as personal factors, environmental factors and therapy provided (Fleuren et al., 2018). If recovery has not started, naturally or as a result of therapy, the person may start in a rehabilitation programme that focuses on compensatory activities for actions of daily living (ADL) (Pandyan et al., 2018). It is important to identify people who could recover after injury, and those who are likely to have poor recovery. This will support the focusing of either a rehabilitation programme to restore normal function or engaging in a compensatory rehabilitation programme. The aim is to improve quality of life and ability to cope with daily function (Pandyan et al., 2018). Therefore, investigating a model that could predict functional independence recovery after a stroke will help to direct physiotherapy/occupational therapy to the best outcome programme in a cost-efficient way.

2.3.1 Recovery of independence

Functional recovery, also called recovery of independence, is defined as the improvement in the ability of a patient to be independent in areas such as self-care and mobility. Functional recovery could be affected by some factors, which would assist and probably have a large influence on the process and extent of this recovery. An example of this is the patient's motivation, ability to learn and family support, as well as the quality and intensity of therapy. According to the International Classification of Human Functioning of the World Health Organization (Giardini et al., 2010), physiotherapists and clinicians are often able to distinguish between the recovery of neurological impairment and recovery of functional independency. Specifically, the restoration of neurological deficits will result in functional recovery. However, functional recovery is not limited to neurological recovery from neurological impairments (Bruce H. Dobkin, 1989). Several studies(Kkel et al., 2004;

Kong and Lee, 2013; Kwakkel, 2009; Simpson and Eng, 2012) indicated that the trend of the functional recovery was non-linear with neurological recovery. It is also steep in the first three months post-stroke. (Yagura et al., 2003).

(Kwakkel et al., 2006) reaches the conclusion that functional recovery does not depend on only restoration of impairment, but also, incorporating compensation strategies. Therefore, it is complex, with many differences, such as spontaneous recovery (natural recovery) and response to treatment in patients. The goal of the rehabilitation process post-stroke is to optimize and increment the changes in recovery. Therefore, it is very important to provide instruments that have the responsiveness to detect and measure changes(Simpson and Eng, 2012).

Functional recovery of the arm involves grasping, holding, and manipulating objects which involves recruitment of various combinations of muscle activity from the shoulder to fingers. In contrast, a minimal amount of recovery of the hemiplegic leg may be sufficient to obtain functional ambulation(Feys et al., 1998).

2.3.2 Mechanism of recovery

Spontaneous neurological recovery is the main pattern of early recovery after stroke and most likely involves partial unknown knowledge of biological processes. This means that spontaneous neurological recovery is insufficiently understood. Biological processes have been identified as playing a role in the neurological recovery following a stroke. In rehabilitation programs, this pattern would be neglected because of the lack of a method capable of measuring the effects of time over the recovery course (Kkel et al., 2004; Kwakkel et al., 1996) .

Several researchers have also suggested that spontaneous recovery of the brain in the first week after stroke likely includes combinations of preservation of the penumbra, physiological and neuroanatomical reorganization, alleviation of diaschisis and reperfusion enhanced by post-stroke angiogenesis damage with compensatory changes extending up to 6 months in more severe strokes (Green, 2003). It would be ideal to identify those individuals who are likely to recover so their maximum recovery potential can be reached and if patients are identified as having poor recovery potential, the focus would be on training the compensatory activities. This would save time, effort and money.

2.4 The International Classification Functions Framework for measurement

The ICF framework describes aspects of a person's health at three levels:

i. the individual body parts and functions,
ii. the individual as a whole (activity) and,
iii. the individual in a social context (participation).

Within the ICF framework, each of these three domains contains different items. The ICF provides specific descriptions that can be used to refer to a specific domain. These descriptions, provided by the ICF framework, are used in this research to guide categorization of the mobility-related deficits post-stroke according to their relevant domains. Detailed discussion regarding the categorization of mobility-related deficits in HD in line with the ICF model is provided below.

i. Body function/Structure

In the context of the ICF framework, body structures can be described as the anatomical parts of the body, whereas body functions are defined as the physiological functions of body systems. For example, muscle strength is seen as a function of the musculoskeletal system, whilst balance is an integrated function of the vestibular, visual, somatosensory and musculoskeletal systems. Muscle strength and balance are linked with the person's ability to move independently. Therefore, balance and muscle strength form the foundation for undertaking a wide range of mobility activities that constitute normal daily life. This includes walking and therefore impairments in muscle strength and balance are known to have negative effects on social activity (participating). Evaluation of body function involves muscle tone testing and movement kinematics characterizing the range of passive and active joint movement. There are many reliable and valid clinical scales for measuring impairments, for example, the modified Fugl-Meyer Assessment of Motor Recovery after Stroke and the National Institutes of Health Stroke Scale (NIHSS).

ii. Activities

This forms the intermediate level of the ICF model. As per the ICF, activity is the component of function which involves execution of a task. Among the most important and common day-to-day activities are tasks that involve mobility components. WHO defines mobility as the "individual's ability to move about effectively in his/her surroundings". In a more general and comprehensive sense, mobility can be defined as the process of moving oneself or changing the position or location of body, or body parts.(Cieza et al., 2009). Scales have been used to measure function, but not a motor pattern, for example the Box and Block Test (BBT)(Mathiowetz et al., 1985). The complexity starts with an explanation of experiments that use the functional tests to evaluate a recovery because that might come from either compensation improvement or results of motor improvements. For this reason it is not possible to distinguish between compensation and motor patterns; in order to overcome this limitation, the Wolf Motor test has been created (Wolfe, 2000).

iii. Participation

This forms the third and last level of the ICF. As per the ICF, participation can be viewed as the involvement in a life situation. Participation restrictions are difficulties that individuals may experience in involvement in life situations (WHO, 2001). Participation may be best described by health-related quality of life measures (Power et al., 1999). Quality of life can be defined as the integration of physical, social and psychological functioning of an individual as being influenced by a disease or therapy (Gotay and Wilson, 1998). It refers to the person's evaluation of their current level of health and functioning as well as satisfaction compared to what they used to have. Buma et al. (2013) highlight the need to distinguish between the neurological recovery at the structure level and the improvement at the activities level. Some studies (Buma et al., 2013; Houwink et al., 2013; Kkel et al., 2004) have reported that maximum recovery can occur during the first three months of a stroke. Also, it may be possible that several motor deficits recover rapidly while other continue to remain as permanent deficits.

In contrast to the lower limb, impairment and disability of the upper limb are more common, and many studies indicate that of the recovery of motor and other functions is also poor and more difficult in the upper limb. Recovery of the upper limb has different patterns of outcomes than the lower limb. For example, upper limb recovery is slower than that of the lower limb. Therefore, patients are more likely to have different rehabilitation needs. Because of the variability that has been seen in each individual's disability after stroke and rehabilitation, outcome measures have been developed to assess and detect change over time or over interventions (Simpson and Eng, 2012).

2.5 Clinical measurement of post-stroke outcomes

The main aim of rehabilitation is to minimise the impact of impairment and maximise the reintegration of the patient who suffered a stroke. However, measuring the effectiveness of interventions is important, both to explain that rehabilitation has occurred and potentially to construct exercises for future management (Barnes et al., 2005).

The assessments of stroke rehabilitation have encouraged the development of many outcomes measures applicable to one or more of its many dimensions. It is broadly agreed that there are three (categories) scales of the individual functioning body – part body. The first scale is used to measure the body structure, the second is utilised to evaluate the activities, and the third scale is used to assess participation. Based on the ICF, there are 38 common assessment tools for stroke patients(Kwakkel et al., 2014). These can be divided as follows:

- 14 tools are used to measure body structure/ functions, the most common are:
    1. Stroke Rehabilitation Assessment of Movement (STREAM).
    2. Glasgow Coma Scale.

3. Fugl –Meyer Assessment of motor recovery after stroke FMA.
4. National Institutes of Health Stroke Scale (NIHSS).

- 15 tools are used to measure activity, the most common are:

  1. Action Research Arm Test (ARAT).
  2. Box and Block Test (BBT).
  3. Wolf Motor Function Test.
  4. Frenchay Activities Index (FAI)
  5. Barthel Index (BI)

- 9 tools are used to assess health-related quality of life outcomes and Participation, the most common are:

  1. Canadian Occupational Measure.
  2. Nottingham Extended Activities of Daily Living (NE-ADL).
  3. Stroke-Specific Quality of Life (SS-QOL).

In this research, the main focus is on the instruments, defined below, that are measures of upper limb motor function. These are selected because they are commonly used in previous studies, such as ARAT(Kwakkel and Kollen, 2007), and have acceptable properties such as reliability and validity (Van Der Lee et al., 2010).

2.6 Measures of upper limb motor function

Outcomes measurement is a result of assessment processes or impairments. It needs to identify the effectiveness of rehabilitation interventions. To be applied in both clinical practice or research, measures must have reliability, validity, and responsiveness to clinically relevant change. Therefore, not only is there a need to be provided with instruments to assess general outcomes, for example Barthel Index, there is also a requirement for instruments to detect changes in rehabilitation intervention in the upper limb. Even though the changes are small, they may be considered essential to the patient or their care givers (Ashford et al., 2008). As a result, several instruments of focal motor function tests have been modified, and these are presented here.

2.6.1 Action research arm test (ARAT)

The Action Research Arm Test (ARAT) instrument has been used to measure the activity of the upper limb (Hsieh et al., 1998). With the patients in a sitting position, a modified table with shelves is brought in front of them and they are asked to perform 19 separate tasks. The ARAT comprises of 19 items divided into four levels (subgroup test): grasp (6 items), grip (4 item), finger pinch (6 items), and gross movement (3 items) of involved upper limb, after Lyle adjusted it in 1981 (Yozbatiran et al., 2008). Each subgroup in the ARAT depends on hierarchical order, the test starts with testing a difficult item followed by an easier item, and after that the items with gradually incrementing difficulty. This means that items are ordered in a sequence, for example, the first item is the most demanding with reference to the level of strength and movement control required, and the second item is the least demanding of sub-group test. An ordinal 4 points scale is scored on to each item. With 0 scores for the patient who could not perform the item, 1 for partial completion, 2 for describing a function which is performed fully, but with abnormal synergies or with difficulty, and 3 for the item which is normally performed. The greatest score of ARAT is 57.

Scores are given according to the different movement and contributions to the overall score of the patient, (Kwakkel et al., 2000). The reliability and validity of ARAT in measuring post-stroke upper limb function has been proven (Yozbatiran et al., 2008). Both intra-rater and inter-rater reliability are reported to be very high with ICC values greater than (0.98), and the test was found to be responsive in detecting the changes during recovery from stroke (Nordin et al., 2014).

ARAT is an instrument that is said to be a more responsive and objective measure of motor activity (Baird et al., 2001). It consists of central properties, showing its function and use. These properties are:

1. It is time-efficient, taking a short amount of time to produce results.
2. It is an easy measure of the upper limb function.
3. It gives an assessment of different tasks over a range of complexity.
4. Most sorts of arm functions are covered by ARAT, involving proximal control and dexterity.
5. ARAT is able to distinguish the abnormality of the movement based on the time it takes to perform and allocate a score of two or three.
6. The ARAT does not need strict conditions of standardisation, such as source, material, weight, and size of tools that are used for testing.

2.6.2 Box and block test (BBT)

BBT is designed as a measure of unilateral gross manual dexterity of patients post stroke. The BBT was developed by Jack (1981) for adults who have cerebral palsy. The BBT was modified and copyrighted in its current form in 1957. The test is simple and quick (Mathiowetz and Weber, 1985). It does not need highly specialised training and requires only simple equipment. It consists of a wooden box that has two equal size parts and fifty equal size blocks placed in the wooden box. It is measured by accounting the number of blocks, one to one, which can be converted by the participant from one part of a box to another part for during seconds. Scores are recorded as blocks per minute for each hand. Higher values mean better gross manual dexterity. Some studies (Hsieh et al., 2009; Platz et al., 2005) have reported the BBT has a test-retest reliability of more than (0.9) and correlates highly with another similar measurement of upper limb dexterity such ARAT. However, as a measurement of upper limb function, the BBT could not afford an assessment of different tasks or ranges. As such, the practice of BBT may be linked to significant floor effects in some patient groups (Mathiowetz and Weber, 1985).

2.6.3 Fugl- Meyer assessment (FM)

The purpose of this measure is to clinically assess the severity of disease, motor recovery and plan of treatment. The Fugl-Meyer Assessment consists of three independent subclasses and that can be used separately or combined into a total motor score. One of which is the upper limb-extremity subscale. It is used to assess the motor impairment of the upper limb for patients in stroke. It consists of 33 items that assess the movement and reflexes of the shoulder, wrist, hand and coordination, with a score out of 66, indicating optimal recovery (Fu et al., 2012; Fugl-Meyer et al., 1975). Each item is scored on a 3- point ordinal scale (0- cannot perform, 1- performs partially, 3- performs fully). It depends on hierarchical order with a ceiling effect (Hsieh et al., 2009).

2.6.4 Wolf Motor function test (WMFT)

The Wolf Motor Test is a common clinical measurement tool used in assessing the patients' motor ability of upper limb post-stroke. It was originally adapted by Wolf et al. (1989), and it was modified by Taub et al. (2011) to measure the influence of power use of upper extremity function (Fritz et al., 2009). It has 17 tasks and begins with placing the hand on a table top that is a simple item. The item's progress is then assessed in a more taxing motor task, such as stacking checkers or picking up a paper clip. The time is limited to a maximum of two minutes, in which all tasks of the test must be complete. A 6-point ordinal scale is used for functional ability, where zero indicates no attempt with the involved arm and five indicates the arm does participate and movement appears to be normal. The test-retest reliability, inter-rater reliability, criterion validity, and construct validity of the WMFT has been ascertained in stroke patients (Fritz et al., 2009; Lin et al., 2009). It is a suitable test for detecting changes over time. This means it has high responsiveness (Hsieh et al., 2009). However, it could not be used to provide information on activity limitations (for example walking and upper limb function) as it is only assesses the level of impairment (Kwah and Diong, 2014).

2.6.5 Motricity index

The Motricity Index is a measure used to assess the deficit of motor movement in a stroke patient. It is used to evaluate the muscle weakness, primarily on the ipsilateral and contralateral sides to the cerebral lesion. It is valid for the upper extremity and is supported by a high degree of relation between its elements and the correlation with both grip strength and a measure of upper limb dexterity function. The Motricity index of each upper limb includes three tasks: pinch grip, elbow flexion and shoulder abduction. For testing the legs, three tasks are also required: ankle dorsiflexion with a foot in a plantarflexed position, knee extension with the foot unsupported and the knee at 90°, and hip flexion with the hip bent at 90° moving the knee towards the chin. These are each scored (0–33) according to the instructions of Collin and Wade (Collin and Wade, 1990). The total upper extremity score involves adding one to the sum of the three actions (maximum possible score=100).

2.7 Outcomes/ dependent variable

The present study is focused on ARAT which is used as an outcome measure for upper limb extremity function after a stroke. This is because ARAT is reported to be the most common measure in the literature. Additionally, it is underpinned by good psychometric properties (Nijland et al., 2013; Stinear, 2010; Stinear et al., 2012) and standardised manner (Yozbatiran et al., 2008). Studies have shown that the ARAT is more responsive to improvement in upper extremity function than the Fugl-Meyer Assessment (FMA) in chronic stroke patients undergoing forced use treatment (Van Der Lee et al., 2001). Furthermore, another study showed that the ARAT was a more stable way of scoring than the Wolf motor test based on a Bland-Altman plot (Nijland et al., 2010a).

2.8 Systemitic literature review methodology

Search strategy: references for this literature review focus on "prediction of upper limb recovery after stroke", reviews published in the English language for humans only, from 1978 to 2015. The search was conducted in MEDLINE with the following keywords: ''predict'', ''forecast'', ''prognosis'', ''upper limb'', ''recovery'', ''stroke'', ''Statistical Models". A search retrieved 369 publications, some of which related to the prediction of stroke; the electronic search method was as the following:

- (S1- S3) were (stroke) or (cerebrovascular disease) or (NH "Ischaemic Attack, Transient") or "Ischaemic Attack, Transient") or "Cerebrovascular*".
- (S4-S11) were (arm) or (hand) or (shoulder) or (elbow) or (wrist) or (finger) or (thumb) or (MH "upper Extremity") or (upper limb).
- (S12-S15) were (predict) or (forecast) or (prognosis) or (MH "Models statistics").
- (S16) was (MH" Recovery of function") or (recover).
- (S17) was (S4 OR S5 OR S6 OR S7 OR S8 OR S9 OR S10 OR S11).
- (S18) was (S12 OR S13 OR S14 OR S15).
- (S19) was ((MH "Infarction, Middle Cerebral Artery") OR (MH "Infarction, Anterior Cerebral Artery") OR (MH "Infarction, Posterior Cerebral Artery") OR "Cerebral Artery").
- (S20) was (S1 OR S2 OR S3 OR S19).
- (S21) was (S16 AND S17 AND S18 AND S20).

- (S22) was (S16 AND S17 AND S18 AND S20) and (Limiters – English Language; Human). More information sees in Table 2-1.

Table 2-1 Process of systematic literature review.

	Keywords	**Method**	**Database**	**No. of Article**
S1	Stroke	Search modes - Boolean/Phrase	Interface- EBSCOhost Research Databases, Database – MEDLINE	208,867
S2	(MH "Ischaemic Attack, Transient") OR "Ischaemic Attack, Transient"	Search modes - Boolean/Phrase	Interface - EBSCOhost Research Databases Search Screen - Advanced Search Database - MEDLINE	17,942
S3	"Cerebrovascular"	Search modes - Boolean/Phrase	Interface - EBSCOhost Research Databases Search Screen - Advanced Search Database – MEDLINE	115,374
S4	Arm	Search modes - Boolean/Phrase	Interface - EBSCOhost Research Databases Search Screen - Advanced Search Database – MEDLINE	285,966
S5	Hand	Search modes - Boolean/Phrase	Interface - EBSCOhost Research Databases Search Screen - Advanced Search Database – MEDLINE	562,183
S6	Shoulder	Search modes - Boolean/Phrase	Interface - EBSCOhost Research Databases Search Screen - Advanced Search Database - MEDLINE	59,746
S7	Elbow	Search modes - Boolean/Phrase	Interface - EBSCOhost Research Databases Search Screen - Advanced Search Database - MEDLINE	29,332
S8	Wrist	Search modes - Boolean/Phrase	Interface - EBSCOhost Research Databases Search Screen - Advanced Search Database - MEDLINE	34,976
S9	Finger	Search modes - Boolean/Phrase	Interface - EBSCOhost Research Databases Search Screen - Advanced Search Database - MEDLINE	129,547
S10	"Thumb"	Search modes - Boolean/Phrase	Interface - EBSCOhost Research Databases Search Screen - Advanced Search Database - MEDLINE	16,600
S11	(MH "Upper Extremity") OR "upper limb"	Search modes - Boolean/Phrase	Interface - EBSCOhost Research Databases Search Screen, Advanced Search Database - MEDLINE	19,924

S12	Forecast	Search modes - Boolean/Phrase	Interface - EBSCOhost Research Databases Search Screen Advanced Search Database - MEDLINE	1,088,066
S13	Predict	Search modes - Boolean/Phrase	Interface - EBSCOhost Research Databases Search Screen - Advanced Search Database – MEDLINE	79,606
S14	"prognosis"	Search modes - Boolean/Phrase	Interface - EBSCOhost Research Databases Search Screen - Advanced Search Database – MEDLINE	525,488
S15	(MH "Models, Statistical")	Search modes - Boolean/Phrase	Interface - EBSCOhost Research Databases Search Screen - Advanced Search Database – MEDLINE	70,046
S16	(MH "Recovery of Function") OR "recover"	Search modes - Boolean/Phrase	Interface - EBSCOhost Research Databases Search Screen - Advanced Search Database – MEDLINE	510,640
S17	S4 OR S5 OR S6 OR S7 OR S8 OR S9 OR S10 OR S11	Search modes - Boolean/Phrase	Interface - EBSCOhost Research Databases Search Screen - Advanced Search Database – MEDLINE	1,017,483
S18	S12 OR S13 OR S14 OR S15	Search modes - Boolean/Phrase	Interface - EBSCOhost Research Databases Search Screen - Advanced Search Database – MEDLINE	1,610,745
S19	(MH "Infarction, Middle Cerebral Artery") OR (MH "Infarction, Anterior Cerebral Artery") OR (MH "Infarction, Posterior Cerebral Artery") OR "Cerebral Artery"	Search modes - Boolean/Phrase	Interface - EBSCOhost Research Databases Search Screen - Advanced Search Database – MEDLINE	29,383
S20	S1 OR S2 OR S3 OR S19	Search modes - Boolean/Phrase	Interface - EBSCOhost Research Databases Search Screen - Advanced Search Database – MEDLINE	315,345
S21	(S1 OR S2 OR S3 OR S19) AND (S16 AND S17 AND S18 AND S20)	Search modes - Boolean/Phrase	Interface - EBSCOhost Research Databases Search Screen - Advanced Search Database – MEDLINE	447
S22	(S1 OR S2 OR S3 OR S19) AND (S16 AND S17 AND S18 AND S20	Limiters - English Language; Human Search modes - Boolean/Phrase	Interface - EBSCOhost Research Databases Search Screen - Advanced Search Database – MEDLINE	369

In total, 369 articles were found. 30 of these studies were excluded, since they were duplicates. 290, were excluded, after reviewing the title and abstract. The forty-eight articles that fitted the search criteria were then included for review and there were only 14 articles that were considered to study prediction modelling of recovery post-stroke. Steps were undertaken by the researcher and the project supervisor to obtain a more robust result. A flow chart shown below presents the search process.

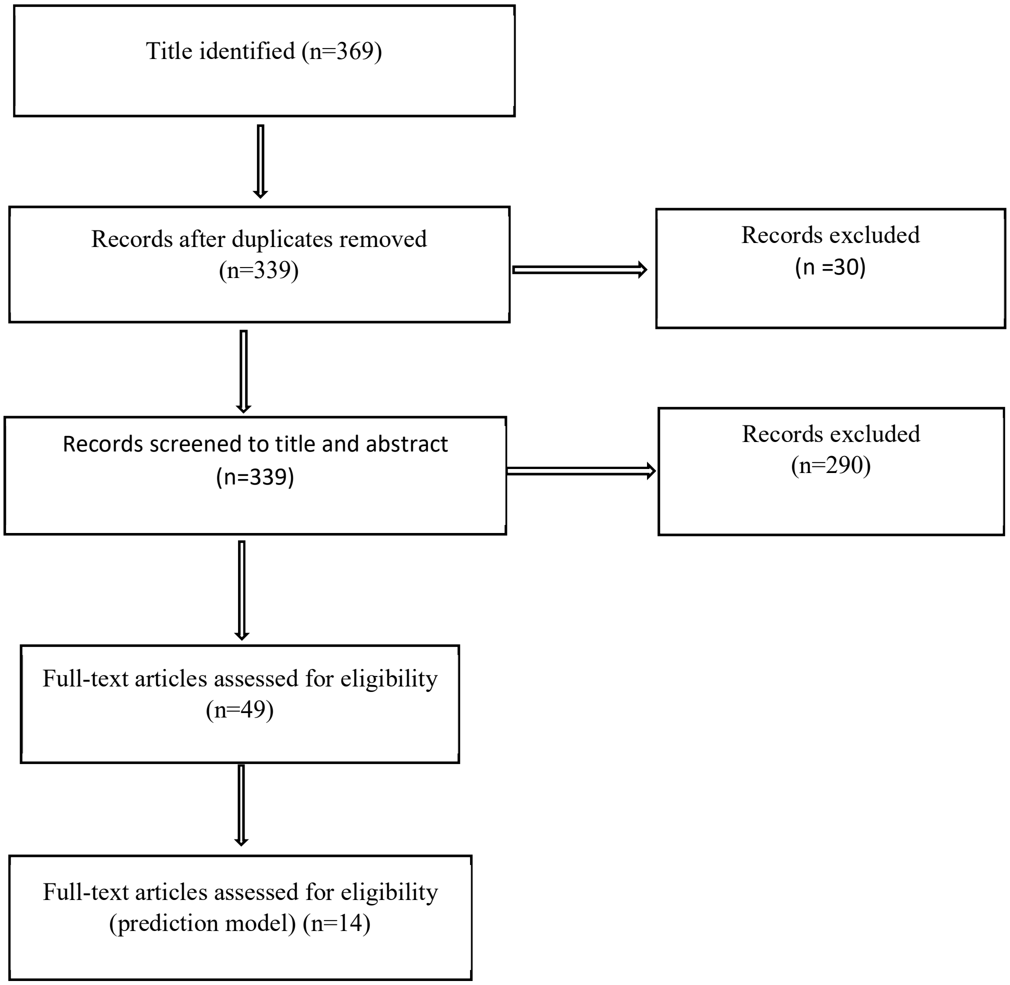

Title identified (n=369)
Records after duplicates removed (n=339)
Records excluded (n =30)
Records screened to title and abstract (n=339)
Records excluded (n=290)
Full-text articles assessed for eligibility (n=49)
Full-text articles assessed for eligibility (prediction model) (n=14)

Figure 2-2 Flow diagram of literature review

2.9 Prediction

Prediction is a term used to define what is expected to happen in the future, or as a definition of problems estimation. The tools that are used to predict are called prediction models. In the clinical field, the prediction model is called a clinical prediction model/clinical prediction rule. The clinical prediction models are clinical instruments that quantify the individual's variables. These variables are analysed and studied to understand their contribution to that individual's diagnosis, prognosis and expected response to treatment. These variables can take many forms, such as an individual's medical history, results from physical examination and other medical investigations (McGinn et al., 2000). These models are used to predict the risk of disease development in a person, or to predict health outcomes in individuals. In stroke recovery, prediction models play a significant role in evidence-based clinical decision-making by objectifying, simplifying and increasing the accuracy of the expected patients' future functioning level (Veerbeek et al., 2011). Thus, there have been differences in studies on predictions made for motor recovery (Feys et al., 2000a), activities (Kwakkel and Kollen, 2013), functional recovery (Wang and Fan, 2014) cognitive function (Suzuki et al., 2013), spontaneous neurological recovery (Arboix et al., 2003), independence in activity daily living (ADL) (Schiemanck et al., 2006; Woldag et al., 2006) and mobility (König et al., 2008). For upper limb recovery, many studies have presented prediction models of upper limb recovery as described in Table 2-2.

Table 2-2 Characteristics of prediction studies that aimed to predict recovery of upper limb post-stroke.

Author, year	***Type of model***	***Outcome/ dependent***	***Variable selection methods***	***Internal validation***				***Discrimination***		***Calibration***	***External validity***
				Apparent validation	***Split sample validation***	***Cross validation***	***Bootstrap validation***	***sensitivity***	***specificity***		
Prsi, 1998	*Logistic regression*	*Hand movement scale*	*Based on goodness of fit*	√	-	-		-	-	-	-
Hilde,2000	*Multiple linear regression*	*Fugel-Meyer*	*R² variable selection and Mallows criteria*	√	-	-	-	-	-	-	-
Alison, 2001	*Multiple Logistic regression*	*Barthel index*	*Both forward and backward*	√	-	-	-	*0.77*	*0.71*	*Used Hosmer Lemeshow*	*Testing with another data set*
Gert, 2003	*Univariate and multivariate logistic regression*	*Action research arm test (ARAT)*	*forward, stepwise approach*	√	-	-	-	*0.74*	*0.83*	-	-
Woldag, 2006	*Multiple linear regression*	*Barthel index and Rivermead Motor Assessment*	*Forward stepwise*	√	-	-	-	-	-	-	-
Nijland, 2010	*Logistic regression*	*Action research arm test (ARAT)*	*Bivariate logistic regression, Forward stepwise*	√	-	-	-	*0.93*	*0.86*	-	√
Veerbeek, 2011	*Systematic Review*	-	-	-	-	-	-	-	-	-	-

Stinear, 2012	Predicting potential recovery algorithm (PREP Algorithm)	Shoulder abduction and figure extension	Using K-means cluster to compare the result.	√	-	-	-	0.88	0.73	-	-
Morris, 2013	Multiple linear regression	Health related quality of life	R^2 variables selection	√	-	-	-	-	-	-	-
Gebruers, 2013	Binary logistic regression	Fugl-Meyer Assessment	-	√	-	-	-	0.84	0.70		
Kwah, 2013	Binary logistic regression	Motor Assessment Scale (MAS)	bootstrap variable selection procedure/ backward stepwise regression	-	-	-	-	Use Area Under the Curve = (0.73 to 0.84)	-	Hosmer-Lemeshow test +Calibration	-
Matsugi,2014	Multiple regression	Functional Independence Measure (FIM)	R^2 variables selection	√	-	-	-	-	-	-	-
Sone,2015	Multiple regression	Manual Function Test (MFT)	stepwise multiple regression	√	-	-	-	-	-	-	-
Persson, 2015	Logistic regression	Modified Ashworth Scale (MAS)	stepwise multiple regression	-	-	-	-	0.91	0.92	-	-

It is important to highlight some points and discuss them relative to the aforementioned studies. Firstly, most of these models have been developed based on the data from restricted sources. This causes a problem because the data, used to develop the models, does not present a typical sample of the broader stroke population. Furthermore, the rehabilitation cohort may have prognoses that are not representative of all stroke patients. Utilisation of such data will be negatively reflected in the developed models. This would render the models to be predictively biased, or at least would limit their predictions to populations characteristically like their own. Consequently, the developed models will not be clinically applicable, except on similar populations (Kwah and Herbert, 2016).

Secondly, even though most of the reviewed studies discussed the technique of their model development (for example multiple linear or binary logistic regression methods), very few studies have illustrated information regarding the probability value used for variable acceptance and methods of related variable selection used in model development. An example is the use of the stepwise methods or some criteria (Bayesian Information Criterion). This reduces the methodological quality of the developed prediction models for upper limb recovery and makes their predictive value clinically less accurate.

Thirdly, most developed models to date cannot specify the expected value of a patient's outcome precisely. Moreover, most prediction methodologies failed to present information about the performance of the developed model and failed to confirm its internal and external validity. This is vital because prediction rules are always less accurate when retested in new/independent patient groups (Kwah and Herbert, 2016).

Finally, the main aim of more accurate prediction is to gain knowledge about various aspects of recovery post-stroke that could be implemented to plan more effective and efficient treatment/therapeutic programs. For example, the selection of a suitable physiotherapy program in order to gain the expected outcome depends on the capacity to decide which stroke survivors are expected to recover the function of the hemiplegic arm (Kwakkel et al., 1996). It is important to a stroke unit management to be able to predict recovery of dexterity and independence in ADL's early enough (within the first 72 hours post-stroke). To achieve this level of clinical accuracy of prediction models, it is important to use a representative population in developing a model, demonstrate appropriate/modern methods of variable selection and test the internal and external validity of the developed model to ensure that their predictive power will be reflected positively into clinical practice.

2.10 Predictors

The upper limb extremity is mildly to severely affected in about 70% of stroke patients (Coupar et al., 2012). Although patients are being treated to improve the upper limb, most of these patients remain with a non-functional affected upper limb. Furthermore, in many cases the improvement in the ability to move the upper limb has been achieved, but the upper limb is not used for daily function (Rand and Eng, 2015). To achieve an efficient model of predicting recovery of upper limb, predictors variables must be easy to collect, reliable and clinically meaningful. The efficient model may be useful to both clinicians and researchers, to explain outcomes, to improve the design and analysis of clinical trials, to determine suitable interventions, and to precisely inform patients of likely outcomes.

Previous studies have investigated numerous variables (Table 2-3) for their ability to predict upper limb recovery. Cioncoloni et al. (2013) have demonstrated the effect of some predictors on the long-term recovery in complex activities of daily living before discharge from the stroke unit. This study reported that the group of predictors such as strength of the paretic upper limb, age, gender, and the ability to perform basic ADL's had a significant effect at 10 days post-stroke and on independence in complex ADL's at six months. Loewen and Anderson (1990) clarified that some rehabilitation variables, such as, Modified Motor Assessment Scale (motor status) and the Barthel Index (ADL's), have the ability to predict the motor and functional outcomes of stroke patients(Loewen and Anderson, 1990). The study of Smania et al. (2007) found that not only was the active finger extension scale a strong early predictor of recovery for independence in ADL's, but also could be essential in order to plan a specific therapy after the onset.

A systematic review of voluntary arm recovery in hemi-paretic stroke was performed to give evidence that the neurophysiological measures and initial sensorimotor abilities are the best predictors of voluntary arm movement after stroke (Chen and Winstein, 2009). It was focused on categorizing the predictive variables and associated outcome measures in terms of International Classification of Functioning, Disability and Health. Steiner's review gave evidence that the review of prediction of motor recovery considered only the predictive value of motor impairment scores, neuroimaging and neurophysiological assessment (Ackerley and Stinear, 2010). Steiner concluded that these tools could be useful in enhancing the accuracy of the final prediction. A systematic review and meta-analysis of predictors of the upper limb post-stroke categorised predictors into five main groups of the predictors as follows, demographic factors, the severity of stroke as a global factor, severity of focal factors, co-factors related to stroke impairment and neurophysiological factors. Furthermore, this systematic review reported that the most powerful predictors of upper limb recovery are the baseline levels of upper limb impairment and function and intact motor- evoked somatosensory potentials (Coupar et al., 2012).

Although there are 85 predictors which have been tested in different ways in previous studies (Gebruers et al., 2014), we need to check the impact of each predictor in the group, as per Table 2-3 below, in order to determine which predictors have a real effect on the prediction to avoid over-fitting and under-fitting in terms of statistical conceptual and which could be developed as a useful model in people who could not be meaningfully measured.

Table 2-3 Predictors of upper limb functional recovery post-stroke.

Demographic and historical predictors	*Clinical measures of impediments*		*Clinical measures of functional activities and measures participants*
	Sensory	*Motor*	
- Age *- Gender* *- Pre-stroke independence.* *- The stroke sides.* *- The lesion size of stroke.*	*- Upper limb sensory* *- NIHSS sensory deficit*	*- NIHSS of arm and leg Motor.*	*- Motricity Index* *- Nottingham EDAL total* *- Barthel Index*

One of the important independent variables is National Institutes of Health Stroke Scale (NIHSS). NIHSS is a measure used to assess the severity of symptoms in patients with cerebral infarcts. The NIHSS was derived from four other scales namely the Toronto Stroke Scale, the Oxbury Initial Severity Scale, the Cincinnati Stroke Scale and the Edinburgh-2 Coma Scale(Brott et al., 1989). The NIHSS includes 15 items, and the total score is between (0- 42). A higher score means a more severe stroke. It is a very simple and quick quantitative measurement. The NIHSS is a reliable, valid and responsive instrument for evaluating the severity of stroke. However, some items of NIHSS have poor reliably for example (level of consciousness, facial palsy, limb ataxia, and dysarthria) (Meyer et al., 2014).

The NIHSS was developed as a clinical stroke assessment instrument. It is widely used to evaluate acute stroke and document the neurological status in stroke patients. It is crucial for predicting the outcome after stroke as it helps physicians to provide accurate information to patients and develop good targets.

2.11 Methods of prediction of recovery

Accurate prediction models have become critical issues when they are used for predicting recovery outcomes of a survivor post-stroke. Many of the reasons for lacking accuracy have been attributed to general factors affecting most potential predicting, such as the selection of predictors, the selection of the statistical estimating method of models' parameters and the increasing lack of validation in stroke's prediction models. It is argued in many studies that these general factors should not lead to prediction deficiency (Murray et al., 2012).

Different statistical models have been employed for predicting recovery (motor, function) post-stroke in different studies (Feys et al., 2000a; Katrak et al., 1998; Kwakkel and Kollen, 2013; Schiemanck et al., 2006; Suzuki et al., 2011). The multiple linear regression has been shown to be the most popular method used in previous studies. This type of modelling has been found to be suitable for predicting the outcome at a fixed time point, for instance, three months post- stroke(Tilling et al., 2001b). However, the functional recovery has nonlinear features over time. Therefore, the linear modelling is not an accurate method for predicting (Koyama et al., 2005).

The study by Tilling et al. (2001a), presented multi-level modelling as a new approach for predicting recovery depending on statistical theory. This study reported that the standard statistical analysis is not suitable for longitudinal outcomes since the number of patients in the study may drop during the time and frequent assessment of the same patients are not independent (Tilling et al., 2001b).

Some studies (Arboix et al., 2003; Cioncoloni et al., 2013; Gebruers et al., 2014; Weimar et al., 2002) depend on logistic regression models that are applied to identify the recovery in patients post stroke. On the other hand, the studies by Suzuki et al. (2006), Koyama et al. (2005) and Gert Kwakkela (2007) determined that spontaneous recovery depends on the progress of time alone. The later concept depends on logarithmic modelling for predicting ADLs in stroke patients by using two measures: FMA and cognitive function soon after stroke. These are taken at two time point assessments which allows plotting of the high fitting curve (Suzuki et al., 2013).

What makes the problem of predicting recovery post-stroke more complicated is the heterogeneity of the patients' outcomes of stroke and the limitations of validating statistical prediction methods. Due to the lack of the three sorts of validation levels being used, the current prediction models have lack of accuracy. Although there are difference studies which have tried to handle and access a simple way of predicting, the prediction models still suffer from the lack of the external validation.

2.12 Discussion

Accurate prediction modelling could have potential to achieve an important role in serving rehabilitation centres or decision-makers in stroke management(Enderby et al., 2017). These tools could help clinicians to deliver patients with more accurate prognoses, clarify goal setting and make a convenient plan for therapies and shorten hospital/centre stay. Ultimately, this accurate prediction could enable efficient utilization of limited stroke care resources. However, there are some limitations that have negatively affected the prediction accuracy of a model. Some of these points have been discussed previously. Firstly, based on the literature, most prediction studies have developed their models on patients who are only involved in rehabilitation or clinical experiments. Consequently, the values produced by prediction models have biases and do not represent the stroke population (i.e. they are more motivated). Clinical experiments, for example, often select patients using strict inclusion criteria and by the nature of rehabilitation studies, the majority of studies are single blinded. In addition, rehabilitation recipients, may have special characteristics depending on the rehabilitation they are receiving, as trials usually have defined rehabilitation protocols. Both conditions will render the sample to be not representative of the wider stroke population (Kwah and Herbert, 2016). Despite that, the models could be useful in assisting clinicians to predict outcomes of patients in rehabilitation.

Secondly, in studies concerning prediction models, outcome measures used as predictors in the models' development vary. This variance could range from using clinical measures to using neuroimaging and neurophysiological tests. Consequently, each model selected different important related predictors. However, none of these studies considered the clinical importance and applicability of the selected predictors in rehabilitation centres/clinics. This means, that important predictors selected by a model could be clinically inefficient to apply, as some are more expensive and clinically very difficult to collect(Counsell et al., 2002; Kwah and Herbert, 2016; Kwakkel and Kollen, 2013).

Finally, several guidelines have been reported that make recommendations about the process of prediction model validation (Altman et al., 2009; Bustamante et al., 2014; Kwakkel and Kollen, 2013; Kwakkel et al., 1996). However, most of the previous prediction models have been developed without checking the model validation. As a result, researchers cannot recommend these models to be implemented in clinical practice and the models developing must be internally and externally validated.

2.13 Conclusion

The literature review introduced various types of studies and pieces of research about prediction of recovery in a patient post- stroke. The outcome of recovery for patients post stroke has heterogeneity and there is no specific technique to measure recovery of function. The recovery of function does not have a linear pattern and the maximum recovery happens in the first three months post- stroke. On average, stroke recovery plateaus three- to six-months post-stroke.

There are three kinds of measurement instruments (clinical measurement, neurophysiological and neuroimaging) used to assess the factors which are related to stroke. Many predictor variables are used to predict the recovery of patients in stroke, for example gender, age, severity of the stroke, limb dysfunction and the location and size of brain lesion. However, it is noted that only a few predictors are able to explain a change in recovery over time or through an intervention.

The most popular statistical models for predicting recovery post-stroke are multiple linear regression and logistic regression. However, based on the literature, these methods do not take enough concern on predictor variables selection, developing and testing the performance of the model, such as internal and external validity for achieving a satisfying clinical prediction (Kwah and Herbert, 2016; Veerbeek et al., 2011). This has resulted in the following limitations: Current models are still misclassifying a certain number of clients or patients. In most of the studies, prediction models of upper limb recovery post-stroke have not been fully tested prospectively. The heterogeneity is so large that some of the models are not representative of an individual. Therefore, the main purpose of this research was to develop a model that can be used to predict an individual's recovery potential using baseline hospital admission data and other demographic variables.

2.14 Highlight points

It seems that developing models is a straightforward process that consists of selecting a modelling approach, linking it with data and producing a prediction model. The method will create a prediction model that might not be as reliable and accurate when using it with a new data set. To produce an accurate model, I was first required to understand the data and identify the model's objectives. Then, I would pre-process and split the data. Only after implementing these steps, did I proceed to developing, evaluating and presenting the models.

Chapter Three

3 Predictors variable selection and models' performance methods

3.1 Regression analysis

Regression analysis is a method of predictive modelling which estimates the relationship between a dependent (outcome) and one or more independent variables (Steyerberg, 2009). Regression analysis is used for finding the causal effect relationship between the variables. For example, the relationship between stroke and age is best studied through regression (Alexopoulos, 2010). It is an important tool for analysing clinical research data.

In its simplest form, regression analysis allows clinician researchers to analyse relationships between one independent and one dependent variable. In medical applications, the dependent variable is usually the outcome we are most concerned with, in this case the recovery from stroke. On the other hand, the independent variables include biographical variables (for example age), neuroimaging variable (for example MRI) and clinical measures (for example, the severity of stroke). The key advantages of using regression analysis are that it can:

1. Explain if predictors have a significant relationship with an outcome.

2. Show the relative strength of different predictors' effects on a dependent variable (outcome) and make predictions.

Since there are numerous metrics of independent and dependent variable and regression line, there are different types of regression styles to make predictions and in this research, I will be using logistic regression, as it is the most common data analysis method for modelling the relationship between a binary response variable/outcome and a set of predictors.

A key part in the regression modelling of data is prediction's variables selection. Over the years, several selection techniques have been proposed in the setting of logistic regression models, and these can be introduced as one case of general linear models (GLMs) cases. Therefore, methods proposed for selecting linear regression models are helpful to exploit approaches in logistic model selection. In fact, some model selection methodologies in logistic regression models are initiated from linear regression(Mille, 2002; Steyerberg, 2009).

Predictors selection is a statistical process which aims to select the best subgroup of predictors and to reduce the redundant predictors in the model. This is an essential step and arguably the hardest part of developing a model, especially with data sets containing many candidate predictors (Ryan, 2008). The idea here is to shrink the multiple/many predictors variables to a smaller subset containing only the paramount variables. The logic behind reducing the number of variables in a model is that the model obtained is more numerically stable and easier to use in practice. When a model is developed without proper predictors variable selections, this could lead to an increase in the estimated standard errors, and an increased dependency of the model on the initial dataset, and therefore overfitting. Overfitting is typically characterized by unrealistically large estimated coefficients and/or estimated standard errors. This can be especially troublesome if the number of model predictors is large relative to the number of sample size.

Predictor selection methods aim to select an optimal subset of predictors variables that contain relevant information, and thereby improving prediction models. This should be achieved by improving the accuracy of prediction and/or simplifying interpretability of the model's results. Additionally, the variance of outcome prediction and parameter estimation is affected by the number of predictors that are chosen. Adding a new predictor would always have an impact on increasing the magnitude of both variances: the model's prediction variance and estimated coefficients' variance.

3.2 Types of methods of predictor selection.

3.2.1 Method 1: Traditional methods

Traditional methods are purely based on statistical significance of the relationship between independent variables and dependent variable. These methods are and continue to be utilised due to their high acceptance rate and popularity among scholars (Ryan, 2008). Although there are many similarities between the model selection in linear and logistic regressions, there are some differences. For example, some criteria of linear regression cannot be applied in logistic regression in the same manner, and vice versa. Traditional methods of predictor selection include: all sub-selection based on the criteria methods and stepwise regression selection (backwards elimination selection methods, forward elimination selection methods, and combination of both).

3.2.1.1 Best subset selection

The best subsets approach aims to find out the best fit model from all possible sub-set regression models. It begins with fitting all models that include one predictor, all models that include two predictors, then three predictors, and so on until the total number of predictors has been completed(Mille, 2002). Then, the subset approach compares all models and selects the best model based on one of the model selections of stopping criteria, which will be discussed in more detail later. Although the best sub-set procedure is straightforward to implement, it does require a challenging computational capacity when the number of candidate predictors (p) is large. If there are p predictors, the number of all possible sub-set is 2^p. As p increases, the number of possible models raises steeply. In general, best subset selection becomes unachievable when the number of predictors is greater than 30. Furthermore, it tends to over-fit a model with irrelevant predictors, and the final model would be very unstable. To overcome this limitation, statisticians developed a method that limits the required computational operations – hence stepwise regression methods(Steyerberg, 2009).

3.2.1.2 Stepwise regression methods

Over the past decades, the most common methods for selecting variables in medical studies are stepwise variants selection methods. These approaches work by including the most significant predictors based on inclusion criteria based on two types of inclusion criteria. The first type includes F-test and T test that are used to test of significance for a set and individual regression coefficients in linear regression. The second is the Wald χ2-test that is used the test of importance for individual regression coefficients in logistic regression(Kutner H.Micheal, 2005). Three types of stepwise subset selection exist. These are: backward elimination, forward selection and a combination of both previous types (Steyerberg, 2009).

3.2.1.3 Backward elimination

Backward elimination method can be used for predictor selection in both linear and logistic regression. It starts with a model that involves all predictors. Predictors are then removed from the model one by one, then removing all predictors with non-significance and then re-testing the model. This is repeated until only predictors with a statistically significant effect on the dependent/response variable remain. The number of models required to fit in backward elimination is equal to $1 + p(p + 1)/2$ models; therefore, it delivers another efficient alternative to the best subset selection method (Ryan, 2008).

3.2.1.4 Forward selection

As a reverse procedure of backward selection, the forward selection begins by testing the significance of effects of all potential predictors, followed by choosing the predictor that had the highest significance level of them. For example, in logistic regression the best fit is a model that has the smallest deviance. Then, the next step consists of sequentially entering the remaining predictors into the model, testing the significance of the added predictor in the model, and finally keeping only predictors that achieve a good model fitting. Finally, the most significant of these candidate predictors are retained to the model (Ryan, 2008).

3.2.1.5 Stepwise regression

The stepwise selection is a combination approach of forward selection and backward elimination. As in forward selection, predictors are included in the model sequentially in a stepwise selection. However, after adding each new predictor, the method may also delete any predictors that become no longer significant at each time a new predictor is added. Such an approach intends to imitate best subset selection while holding the computational advantages of forwarding selection and backward elimination (David W. Hosmer 2013).

3.2.1.5.1 Stopping criteria

The stopping rule for inclusion or exclusion of predictors is the main problem in classical selection methods. It is far more important than the specific variant of the stepwise selection method (for example forward, backward, combined, all possible subsets). Several measures are proposed to help to use the best subset selection, such as Mallows Criteria.

3.2.1.5.2 Mallows criteria

Mallows in 1973 proposed a C_α statistics. It depends on using criterion to compare with a different subset of regression models(Mille, 2002). The criterion includes finding the out-of-sample prediction residual for each model indexed by α; the Mallow's criteria has the formula as:

$$C_\alpha = \frac{\|y - X_\alpha \beta_\alpha\|}{\hat{\sigma}^2} - n + 2p_\alpha \qquad (3.1)$$

Where:

σˆ2 is the unbiased error of the full model. The best model is with a minimum value of criteria. The selection predictors based on these criteria in each step is that the predictor will be selected when it is corresponding the smallest value of criteria; or deleted if it is corresponding the largest value of C_α.

The Mallow's drawback is that it is selecting model with unknown data generating process. For the other types which are Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC).

The best subset selection and stepwise selection methods have the advantage in their availability in commonly used software and their suitability to handle missing data. They are also relatively objective and usually reach their goal of making a model smaller. For example, if another analyst is delivered with the same list of candidate predictors in the same data set, the result of predictors selection would possibly be very similar. This property of stepwise makes it possible to repeat this selection procedure of validation in methods such as Bootstrap method (Steyerberg and Vergouwe, 2014).

However, these methods have many drawbacks, such as instability of the selection. In addition, Steyerberg (2009) reported that the stepwise methods have a lack of stability of the sub-set selection predictors. This means a small change in the data causes a large change in the results, especially their predictive errors; their bias in coefficients' estimation; misspecification of variability and exaggeration of p-values. Ultimately, these drawbacks would worsen provision of predictions' quality than the full model. Additionally, these drawbacks would increase when predictors are correlated among each other, or the model is dealing with a relatively large number of predictors, or both (Frank E harrell 2001). Therefore, the penalized methods play a vital role in selecting predictors and developing models. The next section will review penalized selection methods which are proposed to address the weaknesses of sub-set selection.

3.2.2 Method II: Modern methods

Despite having many drawbacks, the classical methods of best sub-sets selection in predictive models are widely used in practice. In last two decades, a few methods have been suggested to overcome the previously discussed obstacles of the classical sub-set selection methods. These methods are Bootstrapping, Uniform Shrinkage and Penalised Maximum Likelihood that have been developed to improve sub-set selection.

In 1996, Tibshirani presented a new method of selecting predictors that was called least absolute square of shrinkage operators (LASSO). LASSO reduces the predictors that have small coefficients depending on a new penalty for linear regression (Tibshirani, 1996). L1 norm was used instead of L2, and their formula is:

$$Estimated\ of\,(\alpha,\beta) = \arg min \left\{ \sum_{i=1}^{n} (y_i - \alpha - \sum \beta\, x_i)^2 + \lambda \|\beta\|_1 \right\} \tag{3.2}$$

3.2.2.1 Bootstrap of selection

Bootstrap selection method concept is a combination of a bootstrap resampling method and the classical selection variable methods. The idea behind the use of the bootstrap methods is to generate (K) random samples of the data taken with replacement. After which, and for each bootstrap sample, selection predictors methods can be applied. For example, stepwise selection variables methods are used with entry and retention criterion (α=0.05) or predictors are selected from the full model with criterion less than (α), accounted from the Wald- χ2 test and save the result (Efron and Tibshirani, 1994).

In the next step, the selected predictors are ordered and ranked based on the predictors' frequency in all the created bootstrap samples. A threshold criterion is then applied to eliminate predictors from the original model that fitted the original sample; for example, select predictors that repeatedly showed for 50 time. The principle of constructing models, using bootstrap selection (Ryan, 2008), is similar to stepwise selection in the original data, which is dependent on the same stopping rule. For example, predictors with low p-values in the original sample tend to be selected with high frequency in bootstrap samples. Some of these results could improve the model, however, there is no clear evidence of the benefits of this procedure.

3.2.2.2 Regularization methods

The most commonly employed prediction models of recovery post-stroke are classical methods, which is based on typical multivariate linear and logistic regressions (Kwah and Herbert, 2016). In these two regressions, two issues must be addressed when developing a model. The first issue is choosing essential predictors and the second is estimating the model coefficients. However, there are many more modern approaches able to capture higher order interactions in the data for example Penalisation/ Regularisation methods. The methods selected for developing a prediction model in this work were Penalised Logistic Regressions (PLRs). PLRS methods include LASSO, Adaptive LASSO (ALSSO) and Group of LASSO (GLASSO). In this chapter a more detail overview of each of these models is provided and the rationale for choosing these types will be presented.

3.2.2.3 LASSO logistic regression:

I start with the typical logistic regression to describe the LASSO logistic regression technique. Typical logistic regression has been a common approach in clinical prediction studies and clinical research for the past four decades. Logistic regression is a linear classifier that is used when the response/outcome is binary and follows the binomial distribution. Statistically, logistic regression aims to maximize the conditional probability of the outcome given the predictors' information. Let us assume that I have a vector of observations with binary outcomes y_i. Each outcome is associated with p predicting variables that are represented by the design matrix x_{ij} (= the number of patients and j= the number of predictive variables), and the objective is to find the prediction $\widehat{y_i}$, which is calculated using:

$$\hat{y} = \log\big(p(x)\big) = x\hat{\beta} \tag{3.3}$$

Where: $\hat{\beta}$ is the vector of the estimated regression coefficient(s)

The estimation of the unknown coefficients is needed to satisfy the prediction in the (3.2). Then, I use the log likelihood method to estimate these coefficients. The log likelihood is the popular approach for estimating the unknown coefficient(s) and assessing the fitting of the logistic model. The log likelihood can be estimated as follows:

$$p(y|x) = p(x)^{y_i}\big(1 - p(x)\big)^{1-y_i} \tag{3.4}$$

$$p\log\big(L(\beta)\big) = \sum_{i=1}^{n} y_i logp(x) + 1 - y_i \log\big(1 - p(x)\big) \tag{3.5}$$

The idea behind using the maximum likelihood is to find the estimated coefficients of models' parameters that maximize the $\log\big(l(\hat{y}|y)\big)$. I can substitute the $\widehat{y_i}$, as follows:

$$\log\big(L(\beta)\big) = \sum_{i=1}^{n} y_i logp(x) + 1 - y_i \log\big(1 - p(x)\big) \tag{3.6}$$

The process of using the method of typical logistic regression could be limited due to two conditions: first, when the number of predictors is large, and second, the existence of multicollinearity issue among predictors. The high number of predictors and the multicollinearity cause two negative side effects on model performance. Due to the increasing complexity of models dealing with many predictors, the model performs well during the training stage, but the model's accuracy significantly decreases in the testing stage. The second negative effect is the increasing difficulty in interpreting predictors effect with instable estimated coefficients in the model, due to many variables and unstable estimated coefficients.

One method that can counteract this phenomenon is the least absolute square shrinkage of operators. The LASSO word comes from the abbreviated "Least Absolute Shrinkage and Selection Operator". LASSO is the second constrained version of ordinary least square (OLS) method. It was proposed by Tibshirani in (1996) using L1-norm instead of the L2-norm in the first version of penalised methods (Ridge regression). LASSO is in some sense like ridge regression; however, LASSO can give more interpretable results because LASSO can shrink some coefficients to zero. The model's coefficients are bounded by some positive number, hence the penalty. This penalty maximises the log-partial likelihood of the model coefficients (Tibshirani, 1996).

In the context of logistic regression, LASSO refers to the addition of a term to the likelihood function, which is based on the estimated coefficient values. Adding a penalty term to the log-likelihood function due to the typical form of LASSO logistic regression, which could be written as follows:

$$\log\big(l(x_i\beta)|y)\big) = \sum_{i=1}^{n} y_i \log(x_i\beta) + 1 - y_i log\big(1 - (x_i\beta)\big) + \lambda\|\beta\|_1 \qquad (3.7)$$

Where: β refers to a vector of coefficient values and λ is a penalty which controls how strongly penalised the model is. As $\|\beta\|_1$ is Lan 1-norm constraint that is usually chosen to be a positive monotonic function of β, increasing the value of penalty (λ) causes to force all model coefficients (β) to zero. In this situation, any reduction in the negative log likelihood due to predictively useful predictors would be outweighed by the increase due to $\lambda\|\beta\|_1$. Conversely, a value of zero for λ implies no constraint on the model and provides the solution to the ordinary least-squares model.

In the methods of the LASSO family, the issue of identifying the accurate estimation value(s) of the penalisation parameter (λ) is essential and requires to be taken into consideration. The estimated value of the parameters' penalty can have a large impact on the performance of the LASSO family methods. In other words, the penalty plays a vital role in making the variable selection process consistent. In addition, because of its value, it will identify the number of included predictors in the model and the amount of bias term imposed on the estimated regression coefficients (Androulakis et al., 2014; Fan and Tang, 2013). Several methods are considered to estimate the value of this parameter penalty (λ):

(1) The information criteria, which could be Akaike information criterion (AIC) or Bayesian information criterion (BIC), and

(2) Cross-validation (CV), which could be either normal CV or generalised cross-validation (GCV).

Both information criteria and cross-validation will be discussed in more detail next.

3.2.2.3.1 Information criteria

Both of AIC and BIC are very commonly used to determine the selecting tuning parameter value in LASSO family methods. These criteria are extracted from the log-likelihood for the logistic regression model (Gao et al., 2012; Sun et al., 2013), and are noted below:

1. Akaike Information Criterion (AIC) was proposed as the distance between estimated and real outcome in logistic regression models; AIC has the formula:

$$AIC = -2\mathrm{L}(\beta) + 2df(\lambda) \tag{3.8}$$

Where: $L(\beta)$ represents the log-likelihood of logistic regression, λ is the tuning parameter and df is the degree of freedom. One drawback of using AIC is that it causes overfitting in the model's variable selection.

2. In 1978, (Schwarz) proposed Bayesian Information Criterion (BIC), which is considered a more consistent method because it uses strength penalty of the degree of freedom, and has the following formula:

$$BIC = -2\mathrm{L}(\beta) + \log(n)\, df(\lambda) \tag{3.9}$$

Where:

L(β) is the maximum likelihood function of logistic regression, λ is the tuning parameter, df is the degree of freedom and n is a constant that presents the sample size. It is essential to note that the best-estimated value of λ is when it corresponds to the minimum value of these criteria over a grid of λ as in formula (3.8) or (3.9).

3.2.2.3.2 Cross-validation method

Before explaining the cross-validation, it is worthy to mention that the cross-validation was used for two purposes. First, cross-validation is used to estimate the minimum deviance of prediction in the LASSO family methods and for finding/selecting the value of tuning parameter λ. Second, cross-validation is used to test the validation of model performance, which will be discussed further in section 3.5.1(Hastie et al., 2015).

Cross-validation is a technique that divides the studied dataset randomly into k-fold/subsets of equal size. Then, I exclude only one subset randomly and calculate tuning parameters and the mean square deviances of remaining subsets individually. Further, I select the estimated value of a tuning parameter that delivers the smallest deviance of prediction. Finally, I use the excluded subset to test the model's performance. In the penalised logistic regression, for example, the cross-validation is used to find the appropriate penalisation value of parameter λ from the training k-1 folds and holds one-fold for testing the penalised likelihood model. The number of subsamples (k) choice between (5) and (10) (James et al., 2013). Typically, cross-validation can be classified into three types that rely on the sample size, as follows:

- If the sample size is large, we can use more than one-fold for testing of the models' data prediction. In this case, the prediction performance would be evaluated at each value of parameter λ, and the model with the smallest prediction's deviance will be selected.

- When the sample size is medium, k-fold cross- validation will be a convenient approach. Typically, to the K-fold is often be taken between three and ten.
- With a small sample size k will be equal to the sample, the cross-validation is called leave-one-out cross-validation (LOOCV).

Statistically, if I have $u = (xi, yi)$ data-set, then the steps to perform a cross - validation process to calculate the optimal value of tuning parameter can be summarised as follow:

1. Split up the given data set $u = (xi, yi)$ randomly into k equally-sized (u_k).
2. Take one subset out to test the model.
3. Find the estimated coefficient of model parameter β(k) using LASSO family method for each part on the remaining subsets $u_k = (u_1, u_2, \ldots\ldots, u_k)$. I can name $\beta^{\wedge}(k)(\lambda_j)$ of the LASSO estimated coefficients that represent the fitted function $y^{\wedge}{}_{k}(x, \lambda_j)$ of a grid of J values of λ; $j = 1,2, \ldots, J$

Calculate the estimation of the expected prediction error of each estimated model on the folding test sample u_k that is as follows:

$$Prediction\ error_{(k)}(\lambda_j) = \frac{k}{n}\sum_{i=1}^{k}(y_i - y^{\wedge}{}_{k}(x, \lambda_j))^2 \tag{3.10}$$

4. Recalculate both of step in (2) and (3) for all k-fold remaining.
5. Calculate the estimated means square error of k-fold prediction using:

$$k - CV(\lambda_j) = \frac{1}{k}\sum_{K=1}^{k} Prediction\ error_{(k)}(\lambda_j)^{2} \tag{3.11}$$

However, when using the LOOCV technique to estimate the optimal value of the penalty λ, I can utilise one of the following criteria: either select the value λ-min that delivers minimum mean error cross-validated predictors error or use the first λ1se value instead of λmin.

Based on the literature, I can conclude that the cross-validation method can reduce the bias term of the regression sum of square due to splitting the data-set randomly into two parts for training and testing. By contrast, the cross-validation gives a significant and inaccurate result when the sample size is large with a big number of predictors. To solve this problem, Tibshirani presented a new algorithm: Generalized Cross-Validation (GCV) (Tibshirani, 1996).

3.2.2.3.3 **General cross-validation (GCV)**

General cross-validation is a modified version of cross-validation that is used to estimate the tuning parameter of the LASSO family(Efron and Tibshirani, 1994). This method does not need to iterate the refitting model to the different data subsets. The formula of general cross-validation is the validation technique which can resample data by changing the rules of training and testing the samples. It was defined as follow:

$$\text{GCV} = \frac{\sum_{i=1}^{n}(y_i - y^{-i}{}_{i(\lambda)})^2}{n(1-\frac{df(\lambda)}{n})^2} \qquad (3.12)$$

Where: df(λ) is the estimated number of the selected predictor's variables in y^(λ). The best value of λ can be found by minimizing the equation over a grid of λ as:

$$\lambda_{optimal} = argmin\; GCV(\lambda_i) \qquad i = 1,2, \ldots., R \qquad (3.13)$$

3.2.2.4 Adaptive LASSO logistic regression

ALASSO is the new modified version of the LASSO. It was presented by (Zou, 2006) to overcome the inconsistent issue of the LASSO. The LASSO asymptotic setup is somewhat biased, because it forces the coefficients to be equally penalized. To solve this problem, the shrinkage coefficient penalty, which is adapted from the L1-norm penalty, is replaced with a weighted L1-norm penalty. The weighted L1-norm penalty can allow a relatively large amount of penalty for zero coefficients and a small penalty for nonzero coefficients. This process could reduce the bias of the estimated coefficients and improve the variable selection accuracy. ALASSO is an effective process to handle some of the bias in LASSO which could be employed to shrinkage of the estimated coefficients corresponding to essential predictors. Additionally, the LASSO is much more insensitive to many noise covariates.

As previously mentioned, in this project I am focusing on some methods of Lasso family logistic regression model. I assume $y_i \epsilon [0,1]$ is a vector of the binary dependent/ outcome variable, x is a design matrix of p-predictors and β_j is a vector of regression's coefficient parameters, then log-likelihood function is defined as:

$$\ell(\beta) = \sum_{i=1}^{n} y_i logp(x) + 1 - y_i \log\big(1 - p(x)\big) \qquad (3.14)$$

The ALASSO solution is obtained by minimizing the equation as followed:

$$\ell(\beta) = \sum_{i=1}^{n} y_i logp(x) + 1 - y_i \log\big(1 - p(x)\big) + \lambda \sum_{\mathrm{j}=1}^{\mathrm{p}} \mathrm{wj}\left|\mathrm{\beta_j}\right| \qquad (3.15)$$

Where $w_j = (w1, w2, \ldots.., wp)^T$ is vector represent the adaptive weighted penalty that is $w_j = \left(\widehat{\beta_j}\right)^{-\gamma}$. Where $\hat{\beta}$ is an initial penalty that comes from solution of the Ordinary Least Square (OLS) method, LASSO version method or Ridge Regression method(Pan and Shang, 2017). However, using the estimated coefficient of LASSO or Ridge regression, to drive the weighted penalty and applying ALASSO, needs two stages of process:

To find an estimated coefficient of regression using standard LASSO with L1-norm penalty/ ridge regression for the data is calculated to represent as the initial penalty, as follows:

$$w_j = w_j = \left(\widehat{\beta_j}\right)^{-\gamma} \qquad (3.16)$$

Secondly, I substitute the value of the initial penalty, the ALASSO solution is transformed as in the equation bellow:

$$\beta^{\wedge}{}_{ALASSO} = argmin[-\sum_{i=1}^{n}\{y_i ln(x\beta) + (1-y_i)\ln(1-x\beta)\} + \lambda\sum_{j=1}^{p} wj\left|\beta_j\right| \qquad (3.17)$$

3.2.2.5 Group LASSO

As it was explained when discussing their properties, the LASSO and ALASSO of logistic regression have the advantages of delivering simultaneous estimations of model’s parameters and predictors selection. In some cases, the predictor's variables have a natural group structure. Natural group structure means that the variable has more than two categorical levels. For example, severity levels in medical conditions can be divided into mild, moderate and severe, in which case categorical levels of the variable must be converted into dummy variables. Thus, the selection treats an individual variable, which has more than two levels, as a group of variables rather than an individual variable (Yuan and Lin, 2006). From a prediction perspective, one of the most popular tasks is to divide the predictor variables into a different group based on the type of predictor variables. In order to address this type of limitation a new procedure was developed which is called the Group Lasso method for the linear regression model. This method of penalising regression (Ming Yuan, 2006) also can handle the predictors when they are grouped in a linear regression model. The group structure of group lasso is completely known in advance which is a very important property of group lasso compared with another method. Then, in 2008, Meier et al developed the group lasso of logistic regression to overcome the same problems by present a new efficient algorithm that works on penalised regression directly. The group LASSO of logistic regression is defined as:

Suppose that yi is dependent variable with a binary outcome (0,1) and X is a matrix that contains p- dimensional and G predictors. Both types of continuous and discrete (categorical) predictors are allowed. I can code the categorical predictor to be as a group that contains the number of levels of categorical variable mins one, however, a cautious predictor variable contains the only one level. Then I can write the conditional probability logistic regression $p_\beta(x_i) = P_\beta(Y = 1|x_i)$ by:

$$\log\left(\frac{p_\beta(x_i)}{1 - p_\beta(x_i)}\right) = \beta(x_i) \tag{3.18}$$

And

$$\beta(x_i) = \beta_o + \sum_{g=1}^{G} x^T{}_{i,g}\, \beta_g \tag{3.19}$$

Where: β_0 is represented the intercept and β_g is the parameter vector corresponding to the g^th predictors variables. Estimated the vector of parameters is needed. Using the minimizer of convex function to obtain the estimated coefficient of parameters which is solution of group lasso logistic regression. The logistic group Lasso is defined as:

$$\beta_{estimated} = -l(\beta) + \lambda \sum_{g=1}^{G} s(dfg) \left\| \beta_g \right\|_2 \qquad (3.20)$$

Where: L(β) is the log-likelihood function, and λ is the tuning parameter that controls the number of shrinkages or regularisation.

3.3 Model validation

This section focuses on an essential stage after modelling that includes testing the performance of the prediction model. The statistical tools used for testing performance will be presented based on their aims that are classified into generalisation performance, calibration and discrimination of the model. These tools are used not only for testing model performance, but also to make the model's performance of classification simpler to interpret. The model's validation is achieved by testing and comparing the model's performance among developing models.

3.3.1 Model performance assessment

Based on the ARAT, functional recovery level is a discrete variable. However, in this project the ARAT different levels transformed the recovery chance to either 'will probably recover' or 'will not recover'. Therefore, most of the statistical tests introduced in this study are for binary outcomes. Almost all evaluations of dichotomous outcome measures will fundamentally involve interpreting the number of true positives, false positives, true negatives and false negatives (Fawcett, 2006).

Table 3-1shows these measures.

Table 3-1 Shows the general rules to assess the models' performance with a binary outcome.

Metric	General Classification	Predicts patient's recovery	Predicts Patients will not recover
True Positive (TP)	*Right prediction*	*Yes*	*Yes*
False Positive (FP)	*Wrong*	*Yes*	*No*
True Negative (TN)	*prediction*	*No*	*No*
False Negative	*Right prediction*	*No*	*Yes*
(FN)	*Wrong*		
	prediction		

The models' power of binary classifying data can be simply explained by using two concepts: the calibrations and the discrimination of the model(Keidan et al., 1994).

3.3.2 Calibrations

Calibrations refer to how close the predicted outcomes of the model are to the actual outcomes, which means how close the prediction of model equivalent is to the true positive patient's probability of recovery across the range of recovery chances between zero to one(Van Calster et al., 2015). Calibration delivers evidence about the accuracy of the developed prediction model's results when compared to actual results, which only applies to the original datasets.

Calibration was utilised by plotting the graph between the original observations and the estimated probabilities by the model. The model is well-fitting or calibrated if the points distribution on the graph follows a 45 line (Steyerberg and Vergouwe, 2014).

3.3.3 Receiver operator curve (ROC)

The ROC is a universal graphics tool that is used to display the two types of error of all the possible cut-off points. The ROC curve is a plot of test sensitivity as the y-coordinate versus x-coordinate which is represented by 1-specificity or false positive rate, is an effective procedure for assessing the performance of the predictive model. ROC is a conventional method that utilises the simple and easy interpretable plot to assess the 'ability' of a model with binary outcomes(Steyerberg et al., 2010a). The model's ability refers to the model's capacity to discriminate between: (1) the patients who have a chance to recover the UL functioning and all other patients, and (2) patients who are less likely to recover the functional UL compared with all patients. This is achieved by counting the true positive rate, true negatives rate, false positive and false negative rates for every possible point. Table 3-2 is essential to mention that recovery prediction is based on patients' scores in ARAT.

Table 3-2 Contingency table to make decision of binary outcomes.

Test Results	**Recovery patients**	**No recovery patients**	**Total result s**
Recovery Patients	True Positive (TP)	False Positive (FP)	(TP+FP)
No recovery patents	False Negative (FN)	True Negative (TN)	(FN+TN)
Total results			

***Sensitivity and Specificity are Defined as TP/(TP+FN) and TN/(FP+TN) respectively. Positive predictive value and negative predictive value are defined as TP/(TP+FP) and TN/(FN+TN) respectively.

From the contingency table, I would plot the points into XY-coordinates, which enables us to calculate the sensitivity and specificity. Ideally, in prediction models, when the area under the curve equals one, the ROC hugs the top left corner. This is indicative that the model discriminates perfectly between patients who have a chance to recover and patient that do not recover. Nevertheless, when the AUC of a model equals 0.5, then the model performs no better than coincidental results. Additionally, all models will include the point (0,0), which corresponds to predicting a negative outcome for all patients, and the point (1,1), which corresponds to predicting a positive outcome for all patients. When the models improve the ROC, the curve will move away from the straight dashed line toward the top left corner of the plot (which is equivalent to perfect discrimination). This curve is useful for assessing the trade-off between sensitivity and specificity and selecting an operating point for the model being evaluated.

3.3.4 Area under the curve

As previously explained, the importance of both ROC and AUC are completely dependent when testing a model's discriminatory power. ROC curvature depends on AUC score, and vice versa. It is repetitive to discuss AUC after discussing its effect on ROC curvature (more details in the previous section).

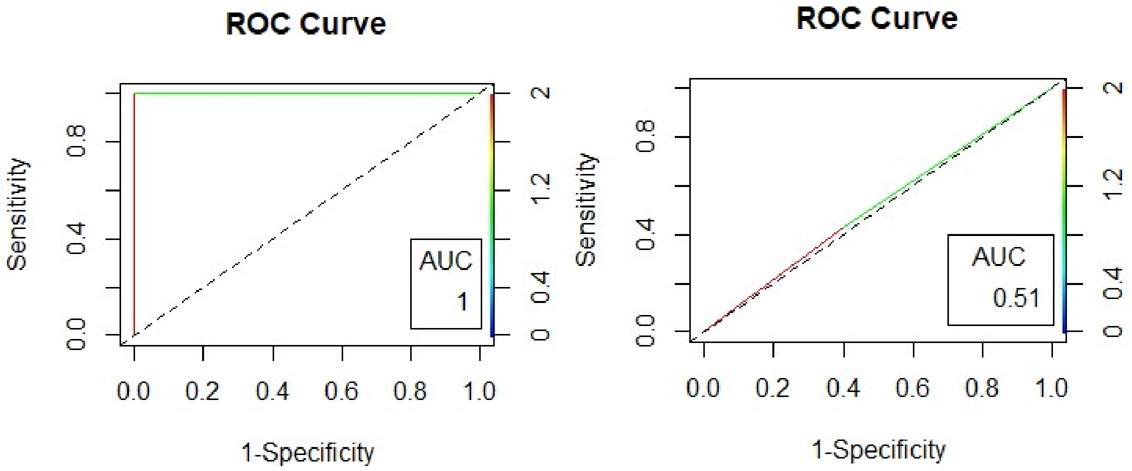


Figure 3-1 In the left plot shows the AUC of perfect model, in the right plot present AUC of deficient model.

3.3.5 Brier score

Brier Score (BS) is the accuracy measure that used to find how close the predicting probabilities are to the actual outcomes using the quadratic score rule(Harrell Jr, 2015). This measure is similar to the coefficient of determination (R^2) in linear regression and has the following formula:

$$BS = \frac{1}{n}\sum_{i=1}^{n}(Y_i - P_i)^2 \qquad (3.21)$$

Where: Y_i is the actual outcome and P represent the probabilities of each patient. The range of score is between (0-1), a score of one means the model's prediction results is inadequate or disagreeing while the score zero indicates that the prediction is perfectly equal to the actual outcomes. However, the middle rate of BS makes the interpretation very complicated to identify model performance that is inaccurate or good. The BS is less complicated than other evaluation measure scores such Nagelkerke's R^2.

3.3.6 Log-likelihood function

The log likelihood is a measure commonly used to evaluate the fit of the model. For a binary outcome which has the binomial distribution, the log likelihood can be evaluated as follows:

$$\log\big(l(p;y)\big) = \sum_{i=1}^{n}(y_i \log(p_i) + (1-y_i)\log(1-p_i)) \tag{3.22}$$

Here it is essential to state that the likelihood improvement is an advance on the log-likelihood function of the model when using a set of predictions, p, against a null model which uses the mean of the outcome as the prediction for observation. The likelihood improvement is calculated as:

$$likelihood\ improvement = \frac{log\big(l(outcome's\ mean; y)\big) - log(l(p;y))}{log(l(outcome's\ mean;y))} \tag{3.23}$$

Deviance

Deviance is used to evaluate the goodness of fit of a logistic regression model. Deviance plays the same role of sum square error (SSE) in the linear regression model (Harrell and Lee, 1984). It compares between observed values of the outcome (response variable) and predicted values comes from models. It was derived based on the log-likelihood function as in the formula:

$$\log\big(L(\beta)\big) = \sum_{i=1}^{n} y_i logp(x) + 1 - y_i \log\big(1 - p(x)\big) \tag{3.24}$$

And then the comparison process of the likelihood is deduced by finding the proportion between the likelihood of the fitted model over the likelihood of the saturated model, as follows:

$$D = -1ln\left\{\frac{likelihood\ of\ the\ fitted\ model}{likelihood\ of\ the\ saturated\ model}\right\} \tag{3.25}$$

Using the equation (3.22) and (3.23) above becomes:

$$Deviance = -2\sum\left[yiln(\frac{p(x)}{yi} + (1 - yi)\ln(\frac{1 - p(x)}{1 - p(x)})\right] \tag{3.26}$$

Where: $p(xi)$ is the predicted values of the outcome. Saturated model refers to a model that contains as many predictors' parameters as there are data a point

3.3.7 Hosmer- Lemeshow test

The Hosmer–Lemeshow is a useful test of the predictive values/ probabilities of binary outcome models by testing the model versus the assumption of correctly calibrated (David W. Hosmer 2013). In this test, the predicted values of outcomes are calculated based on the estimated parameters of the model for each observation in the sample using the equation as follow:

$$p((y = 1|x) = \frac{e^{X\beta}}{1 + e^{X\beta}} \tag{3.27}$$

X is a matrix that represents the predictor's variables, β is a vector representing the estimated coefficients regression and y is a vector that represents the outcomes.

The observation is then divided into ten groups (deciles) based on the predicted probabilities. This mean, and each part's predicted positive outcome proportion is compared with the parts' observed outcome proportion. The χ2 test is used to check the range of difference between the predicting outcome and the perfect fit by approximating the sum of the range of deviations with a χ2 distribution, the formula for of 3.4.6 Hosmer- Lemeshow Test is as follows:

The test examines how well the percentage of patients who have recovered functional upper limb matches the rate of predicted patient's recovery rate deciles of predicted rate.

3.4 External validation

External validation is a process to explore the substantial differences between the characteristics of the two sources of dataset, for example, between the development and validation datasets and to test how well the model performs (Collins et al., 2014). Because of the optimism problem (overfitting) of predictive models, this leads to models having worse performance in new patients/subject than expected from results based on the performance estimated from the development data-set(Harrell et al., 1996; Kwah and Herbert, 2016). Therefore, a process of external validation is considered an essential stage after developing a model to support the general model's applicability in clinical practice (Steyerberg, 2009).

3.5 Evaluation of generalisation performance

There are several statistical methods for testing the generalisation performance of any prediction model based on a new set of data from an underlying population. But some of these techniques can check the model efficacy based on splitting and resampling using the same data and delivers an optimistic estimate of model performance (Efron and Tibshirani, 1994). These methods are suggested to improve the model performance and to avoid the overfitting problem, especially in complex models. The overfitting problem usually happens with complex models. For example, the complex models can gain the ability to perform well based on the training set and testing on a specific set. However, these models, which have a perfect performance in developing based on a set of data, will have a low level of performance and fail with a new data set (external dataset). This weakness is caused by the inflation in the variance of the model performance with the new dataset. At this part of this chapter, it will be mention on the splitting and resampling methods.

3.5.1 Cross-validation

Cross-validation includes splitting the dataset under study randomly into a subset of equal size, to assess the validity performance of model development. It works by dividing the data set into k-fords, holding one out and developing a model for each reaming part; this process is called a model training stage. The holding out is used to testing the model performance. This is an advantage because more than 80% or 90% are used in this stage, however, in other method, for example half splitting that used half of data for training and another half for testing(Trevor Hastie, 2015).

For estimating the test error associated with of penalised methods, CV was used to evaluate penalised methods performance or to select the appropriate value of penalty. An advantage of these method is it is a simple process to estimate the mean square errors.

3.5.2 Bootstrap method

The bootstrap method represents the technique of sampling that aims to find the empirical distribution of the sample of the study. Bootstraps samples are drawn and the model is tested by estimating the calibration and discrimination based on these samples (Efron and Tibshirani, 1994). The coefficients of the model were used without refitting the model, so it would not allow the coefficient change. Statistically, each observation in the original sample has the same chance/probability to select (pi= 1/n), n is the number of patients or rows in the dataset. I then draw samples from the original sample, equal to sampling from the original data with replacement. The model is developed on each bootstrap sample. By contrast, an observation which is not selected in the bootstrap sample has a probability e-1 that can be accounted as follows:

$$pr(not\ selected) = (1 - pi)^n \tag{3.28}$$

$$(1 - pi)^n = (1 - \frac{1}{n})^{\ n} \tag{3.29}$$

$$= (\frac{n-1}{n})^n \approx e^{-1} \tag{3.30}$$

Predictions on this coincidentally held out set predictions have been observed to be unbiased. The approximately equals sign is due to the possible non-uniqueness of each observation, even though this is very unlikely when data has several multiple continuous predictors(Breiman, 2001).

3.6 Decision curve analysis

The prediction models in this research aim to classify of expectations of patients' recovery to recovery and no recovery, which could help to guide rehabilitation programs of patients with upper limb impairments. Therefore, a cut-off point is required to classify patients as either not likely to recover (no treatment) or likely to recover (treatment is indicated). As mentioned in chapter two, the cut-off point is a decision threshold based on the patient outcome, for example ARAT outcome. At the threshold, the likelihood of improvement exactly balances the likelihood of no recovery e.g. improves the clinical costs-effectiveness. In spite of the fact that prediction model may achieve a good level of calibration and discrimination (sensitivity, specificity and the area under the curve of ROC), these characteristics do not enable the model to assess clinical usefulness (Steyerberg and Vergouwe, 2014; Zhang et al., 2018).

To overcome this weaknesses, Vickers and Elkin (2006) have proposed decision-analytic measures to summarize the performance of the model in supporting decision making. Additionally, they derived a new tool as a part of decision curve analysis (DCA) based on subtracting the rate of all patients who are false positive from the rate of true positive. Then, the subtraction result was weighted by using the relation between the false-positive and false-negative results of a prediction model. This tool is called a Net Benefit (NB) that refers to weight a relative between the two false conditions have a formula as follows:

$$Net\ Benefit = \frac{TurePostiveCount}{n} - \frac{FalsePositiveCount}{n}(\frac{p_t}{1-p_t}) \qquad (3.31)$$

Where:

- True- positive count and false- positive count represents the number of patients with the true and false positive prediction models results.
- n is the sample size (total number of patients).
- p_t: is where the expected benefit of intervention is equal to the expected benefit of avoiding intervention.

There are two important benefits behind using DCA. First, DCA can be used to compare different types of models. For example, compare results from a predictive model and results from the clinical decision. Secondly, it can be easy to quantify the prediction models' benefit in clinical practice in a simple way that does not require information on the cost-effectiveness' or how patients perceive their different health states (Holmberg and Vickers, 2013; Van Calster et al., 2018).

3.7 Variance inflation factors (VIF)

Multicollinearity refers to the existence of correlation between the predictor's variable in the model which always causes the inflation of the variance of estimated parameters in the multiple linear regression models. The VIF is a scale that used to detect a multicollinearity level in the model(Steyerberg, 2009). To evaluate by applying the formula as follow as,

$$VIF = \frac{1}{1 - R^2(i)} \tag{3.32}$$

Where:

$R^2(i)$ is the R^2 value that the result from the predicting xi on the other predictors in the regression model. When VIF equals one this means that the correlation between the predictor and the remaining predictor's variables equal zero. If VIF locates between four and less than ten, it means that there is a low level of multicollinearity, while VIF exceeding ten is a sign of severe multicollinearity requiring correction.

3.8 Cluster analysis:

Due to the heterogeneity of stroke recovery outcome, some different approaches might be better suited to overcome this issue. One of these approaches is cluster analysis. Cluster analysis is a useful multivariate method that aims to classify a sample of subjects (or objects) on the basis of a collection of measured variables into a number of variety class such that similar cases are placed in the same group, that is, homogenous, but are very dissimilar to objects in other clusters, that is, heterogeneous (Aggarwal and Reddy, 2013). The two most widely employed techniques for clustering are presenting, as follows:

3.8.1 Hierarchical clustering:

The technique of clustering depends on the idea that it finds a nested sequence of clustering. Two different ways have been employed to achieve clustering, namely, divisive (bottom-up) agglomerative or (top up) clustering. The divisive way includes four steps which are: assign each point of data to single cluster, compute the similarity between each of the clusters and then dividing the cluster to two least similar clusters. Finally, repeat step two and three until there is no single cluster left. While the agglomerative way is the opposite of the divisive way. Both ways utilise the concept of dendrogram which is defined as the development of binary tree based on data structures, see Figure 3-2

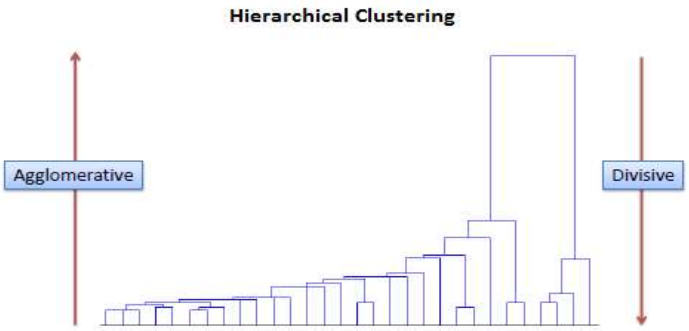


*Figure 3-2 Path of two algorithms of clustering (Divisive and Agglomerative)(*Sayad, 2010-2019*).*

The hierarchical clustering requires accounting the proximity metric which represents the distance between each cluster. Three methods have been used to measure the proximity matrix which is a single linkage, complete linkage and average linkage. Hierarchical clustering does not require the number of clustering and is easy to implement (Clarke et al., 2009).

3.8.2 Non-hierarchical methods or partitioning method

Partitioning methods typically need the number of clusters and initial seeds (or clusters) as an input to the methods. The clusters are then iteratively improved. They try to determine all cluster optimally in one step. The K-means and K-median are the most common partitional clustering.

3.8.2.1 K-means clustering:

K-mean method is the most widely employed partitional clustering (Tibshirani et al., 2001). It requires the number of clusters (k) and the initial centres, one for each cluster. It aims to minimise the square of the distance between each point within the cluster and the position ofμ_i. It aims to minimize the sum square error (SSR) score for the given set of centroids(Aggarwal and Reddy, 2013).

$$sse(c) = \sum_{k=1}^{K} \sum_{xi \in C_k} \|x_{i-}\mu_k\|^2 2 \qquad (3.33)$$

Where x_i represent the dataset and $\mu_k$is the centroid of clusters C_k.

3.8.2.2 K-medians clustering:

K-medians method aims to use the median of each cluster rather than mean of the cluster. K-median clustering select K cluster centres by minimizing the sum of the distance between each point and the closet cluster centre. The distance measure used the L1 norm as opposed to the measure of the k-means and the absolute error rather than a square error. K-median is more robust in handling outliers than k-mans(Clarke et al., 2009). However, like all methods of centroid it works best if the clusters are convex. The function of objective k-median is:

$$S = \sum_{k=1}^{K} \sum_{xi \in k} \left| x_{ij-ME \quad kj} \right|^2 \qquad (3.34)$$

Where: x_{ij} is the sample data and MED_{kj} is median of the data.

There are two factors affecting the performance of partitioning clustering methods(Chen et al., 2002):

1. Selecting the initial centroid.
2. Estimating the number of clustering.

Several methods have been proposed to determine to each of these factors. I describe the K-mean++ and Silhouette method as follows:

3.8.2.3 K-means++:

K-means++ was identified by Arthur and Vassilvitskii (2007) that selecting the centres c1 which is chosen uniformly at random from data set. After that, new centres ci selected $x \in X$ with probability as follows:

$$pr = \frac{(data)^2}{\sum_{x \in X}(data)^2} \tag{3.35}$$

Finally, repeat these steps until it has been taken k centres altogether.

3.8.3 Silhouette method:

Silhouette has been used to assess clustering result by studying separation distances among results. The measure is a range between [-1, 1]. One or closed on value indicates that i is well-matched to its own cluster, and poorly-matched to neighbouring clusters. If most points have a high silhouette value, then the clustering solution is appropriate. In contrast, if many points have a low or negative silhouette value, then the clustering solution may have either too many or too few clusters. The silhouette clustering evaluation criterion can be used with any distance metric(Chen et al., 2002; Rousseeuw, 1987). Silhouette value for i the point, s_i, is:

$$s_i = \frac{bi - ai}{\max(ai, bi)} \tag{3.36}$$

Where: a_i is the average distance from the ith point to the other points in the same cluster as i, and b_i is the minimum average distance from the ith point to points in a differed cluster, minimized over clusters.

3.9 R software packages

R software contains many packages that use variety forms of regression modelling. One group of these packages is concerned with fits generalised regression, cluster analysis and penalised logistic regression problem that imposing a constraint on parameters, for estimating of the entire ridge, LASSO, adaptive LASSO and group of LASSO. Additionally, several cross-validation routines allow optimisation of the tuning parameters.

Here, I introduce brief information about the main R packages of penalised methods, clustering analysis and model performance that have been used in this research.

Package Name	Description	Properties	Tuning parameters
glmnet	It is an efficient process that used to fit the penalised methods (LASSO and ridge regression) of logistic regression models (Friedman et al.).	Ridge and LASSO model of linear, logistic, multinomial, Cox and Poisson models. Cross-Validation with K fold to find the optimal tuning parameter.	Lambda, Alpha
grplasso	Methods that used to fit the penalisation with group LASSO general linear model based on the (Meier et al., 2008)	Fitting of a group LASSO of linear, logistic and Poisson methods	Lambda
parcor	Includes Algorithms for accounting the partial correlations matrix using different types of penalisation methods. It delivers cross-validation model selection for four methods of LASSO family as Well (Zou, 2006).	Four penalised regression methods for the estimation of partial correlations: LASSO, adaptive LASSO, ridge regression, and Partial Least Squares.	-

penalised	An efficient technique for fitting the LASSO or elastic-net regularisation path for some GLM (Goeman et al., 2012)	Elastic net methods paths for linear, logistic, Poisson and Cox models; k-fold cross-validation for optimal lambda1 and lambda2; positivity constraint on regression coefficients	lambda1, lambda2

Packages name	Description	**Propose**
pROC	Provides algorithm for accounting the receiver operating characteristic (ROC), Area under the curve (AUC) and the confidence interval of AUC*(*Sun and Xu, 2014*)*	To visualise and compare the model performance.
ROCR	Contains some flexible function for plotting sensitivity/specificity curves. In addition to, curves come from cross-validation or bootstrapping runs can be averaged and standard deviation or box-plot *(*Sing et al., 2004*).*	ROC graph and creating cut-off parameterised 2D performance curves.
AUC	Contains functions to account the area under the curve of selection measure.	To compute different types of the area under the sensitivity curve, specificity curve, the accuracy curve and the area under the receiver operating curve (AUROC)
rms	Includes several functions that work with many types of regression models, especially with logistic regression models*(Harrell Jr, 2015).*	For the estimation, testing, prediction, and validation of the regression models.

Packages name	**Discerption**	**Propose**
Cluster	Methods of grouping data that extended based on the original form of Peter Roussseeuw.	Hierarchical clustering and Partitioning methods.

Factoextra	Includes simple some functions to extract and present the output of multivariate data analysis, for example cluster analysis*(*Kassambara, 2017*).*	Simplifying a part of the process of clustering analysis and delivers functions of plotting in elegant data visualisation.

Chapter four

4 Modifying cut-off point

The aim from this chapter is to determine a cut-off point that used to dichotomised Action Research Arm Test (ARAT) using cluster analysis method and bar-chart plot. Additionally, to understand the trajectory/pattern of patient's recovery over three months.

4.1 Method of identifying ARAT cut-off point

At first, I adopted data from a secondary anonymised dataset with 178 patients with 300 variables reported by Church et al. (2006). The dataset includes: 1) the Action Research Arm Test (ARAT) outcomes to measure the upper limb function in three different times (baseline: 0-1 weeks post-stroke, second: after 4- weeks intervention and third: at three months post-stroke); and 2) other measures that are used to assess the motor and function of upper limbs status of patients, too. These measures are, for example, Frenchay Arm Test, Motricity Index, handedness, new neurological impairment National Institutes of Health Stroke Scale.

Secondly, I extract a group of patients from the RCTs, more details in next section. Then, I aimed to identify, if possible, a new cut-off of Action Research Arm Test that is clinically meaningful in practice. The cut-off refers to separating two differing statuses, for example separating patients into two categories: recovered and not recovered patients. This binary result (0,1) will enable us to utilise logistic regression analysis. Logistic regression is influenced by the cut-off point that used to classify the outcome of the ARAT score as zero or one. The zero represents a patient that had no recovery of the upper limb, and one represents a patient who recovered. In this project, I studied the patients who had ARAT scores of less than ten for two reasons. First, the groups of patients who have ARAT scores of more than ten have a big chance of recovery, but the patients having ARAT scores of less than ten do not have a clear upper limb recovery based on the ARAT score. The second reason is to try to determine a better cut-off point than the cut-off point in literature for categorising the ARAT scores as binary outcomes. The cut-off points based on literature are one and ten (Kwakkel et al., 2003; Nijland et al., 2010c).

Hierarchical clustering was applied to statistically investigate the new cut-off point value. Additionally, the hierarchy was deduced based on the ARAT subgroup at baseline patients' scores. This was performed to cluster patients to: 1) the patients who performed only gross movements items, or 2) who performed any hand function that is included as an item of the ARAT.

4.2 Identifying the cut-off point:

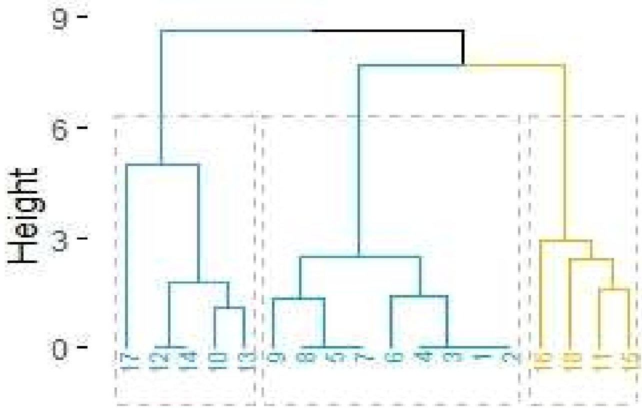
The ARAT scores of eighteen patients were included for determining a cut-off point. All those patients have the total score of ARAT less than 10. Furthermore, the total score of baselines NIHSS outcome and ARAT after three months of those patients are included; see more detail in Table 4-1. The idea behind this table is to introduce evidence that 55% of this group of patients have full functionality of the upper limb recovered after three months based on the ARAT score. In spite of a group of patients having the score of (ARAT<=9) and scores of severities (NIHSS >= 9), some patients went on and recovered. This could support our aim to identify a new cut-off point of ARAT score.

Table 4-1 Patients who have ARAT score less than ten at baseline and outcome of NIHSS.

NO.	*Grasp*	*Grip*	*Pinch*	*Gross-movements*	*Total of ARAT*	*Total ARAT (3 months)*	*NIHSS*
1	0	0	0	3	3	57	2-8
2	0	0	0	3	3	57	>=9
3	0	0	0	3	3	57	>=9
4	0	0	0	3	3	26	>=9

5	0	0	0	4	4	6	2-8
6	1	0	0	3	4	57	2-8
7	0	0	0	4	4	54	2-8
8	0	0	0	4	4	19	>=9
9	0	1	0	4	5	57	>=9
10	0	0	0	5	5	57	2-8
11	2	0	0	4	6	57	>=9
12	0	0	0	6	6	57	>=9
13	1	0	0	5	6	45	>=9
14	0	0	0	6	6	57	2-8
15	3	0	0	3	6	43	2-8
16	3	2	0	3	8	36	>=9
17	0	0	0	9	9	39	>=9
18	4	0	1	4	9	57	>=9

The hierarchical clustering was used to group the 18 patients based on the sub-score group of the ARAT as it is shown below in Figure 2-2.



Number of clusters with the patients' actual number

Figure 4-1 Hierarchical clustering result of ARAT scores less than 9; X- axis represents the cases is grouped. For example, number 11 corresponding to patient 11 in the Table 4-1. The Y- axis represents the distance or dissimilarity between clusters.

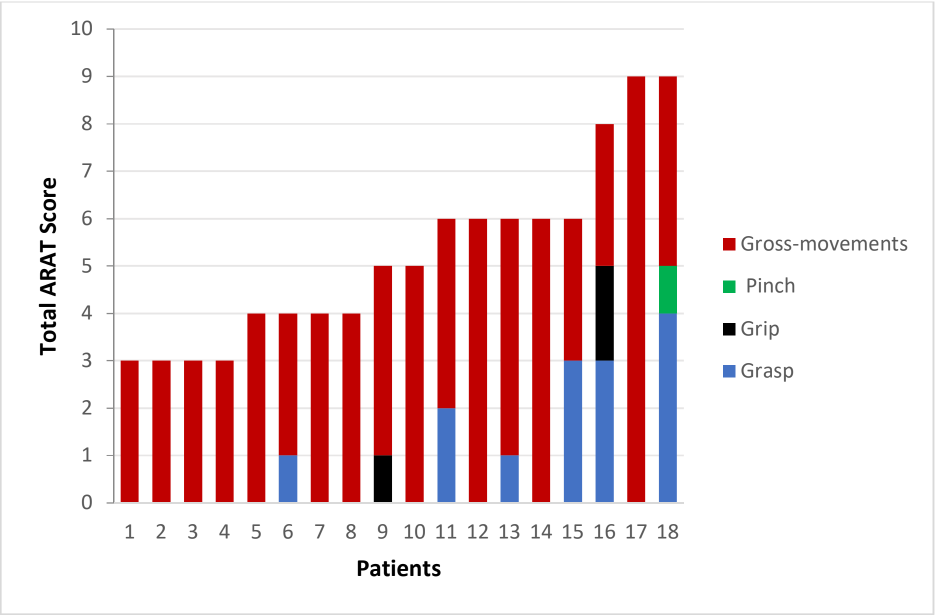


Figure 4-2 The details of ARAT score for patients with a score less than ten. The total score of each patient is plotted on the x-axis and the total score of ARAT is on the y-axis.

To validate the seven values as a cut-off, I took only patients who have ARAT score of less than ten at baseline. The baseline score of ARAT includes the score of each item in the test, the total of sub-group and the total of the ARAT scores. Then, bar-chart plotted ARAT subgroup score (grasp, grip, pinch and gross movement) of patients who have the same total scores Figure 4-2. Eighteen patients were classified to: four patients have total score three, Figure 4-3, four patients have total score four, Figure 4-4, two patients have total scores five Figure 4-5, five patients have scores six

Figure 4-6, one patient has the score eight Figure 4-7, and finally, two patients have scored nine Figure 4-8.

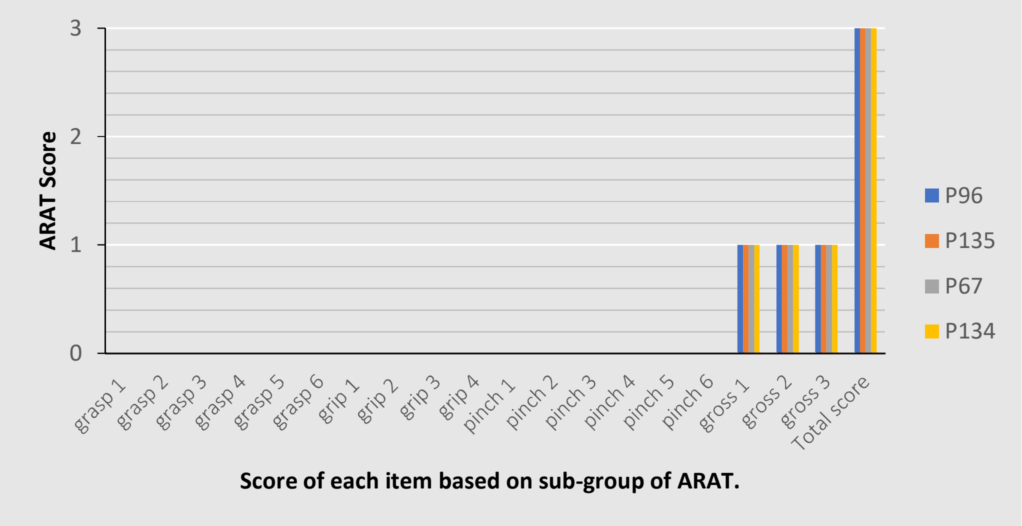


Figure 4-3 Patient with Action Research Arm Test (ARAT) of three. The score of each item based on sub-group of ARAT is plotted on the x-axis and the total score of ARAT is on the y-axis.

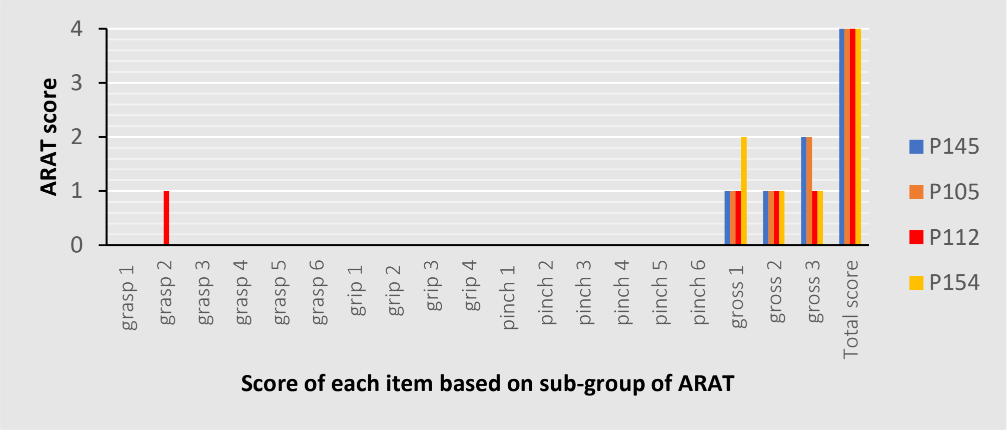


Figure 4-4 Patient with Action Research Arm Test (ARAT) of four. The score of each item based on sub-group of ARAT is plotted on the x-axis and the total score of ARAT is on the y-axis.

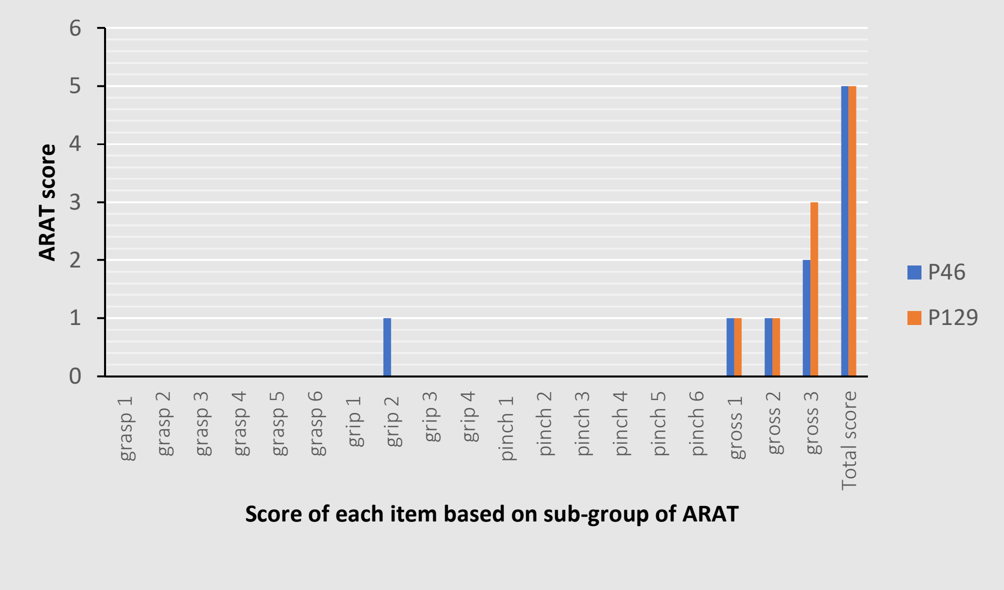


Figure 4-5 Patient with Action Research Arm Test of five. The score of each item based on sub-group of ARAT is plotted on the x-axis and the total score of ARAT is on the y-axis.

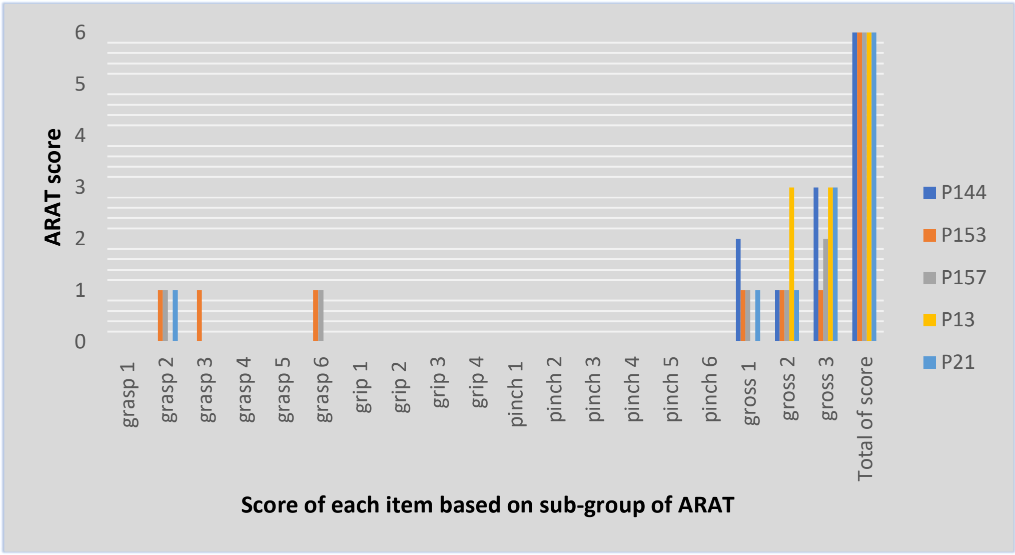


Figure 4-6 Patient with Action Research Arm Test (ARAT) of six. The score of each item based on sub-group of ARAT is plotted on the x-axis and the total score of ARAT is on the y-axis.

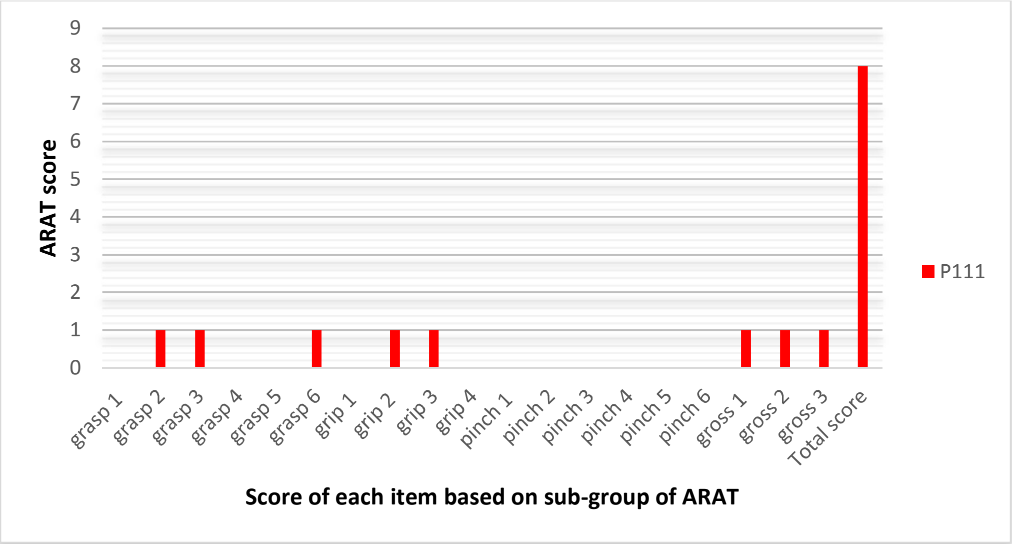


Figure 4-7 Patient with Action Research Arm Test (ARAT) of eight. The score of each item based on sub-group of ARAT is plotted on the x-axis and the total score of ARAT is on the y-axis.

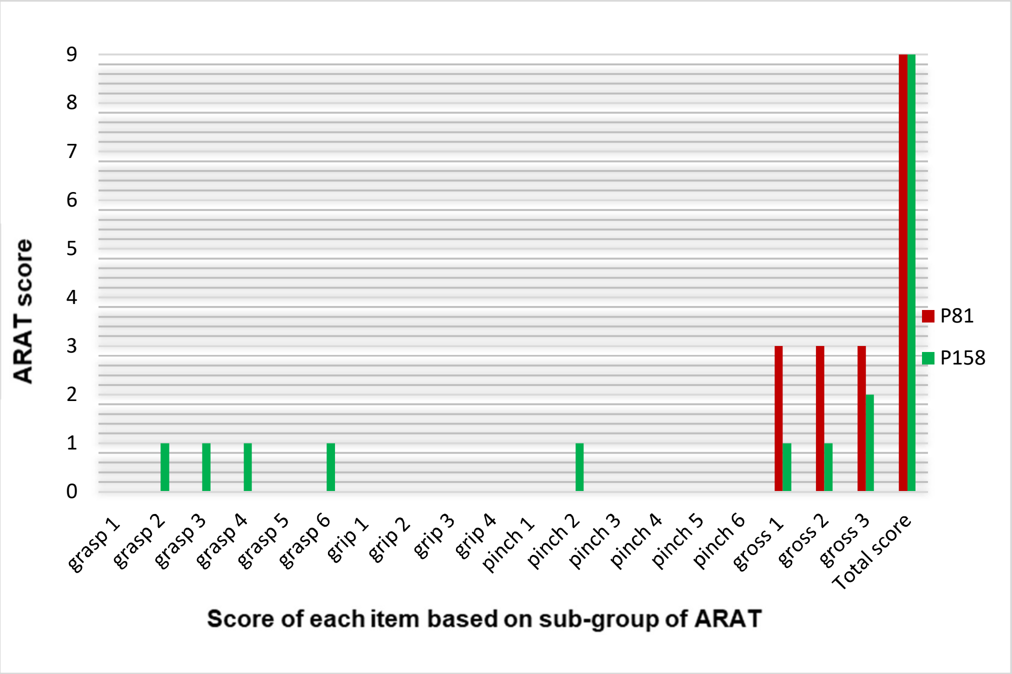


Figure 4-8 Patient with Action Research Arm Test (ARAT) of nine. The score of each item based on sub-group of ARAT is plotted on the x-axis and the total score of ARAT is on the y-axis.

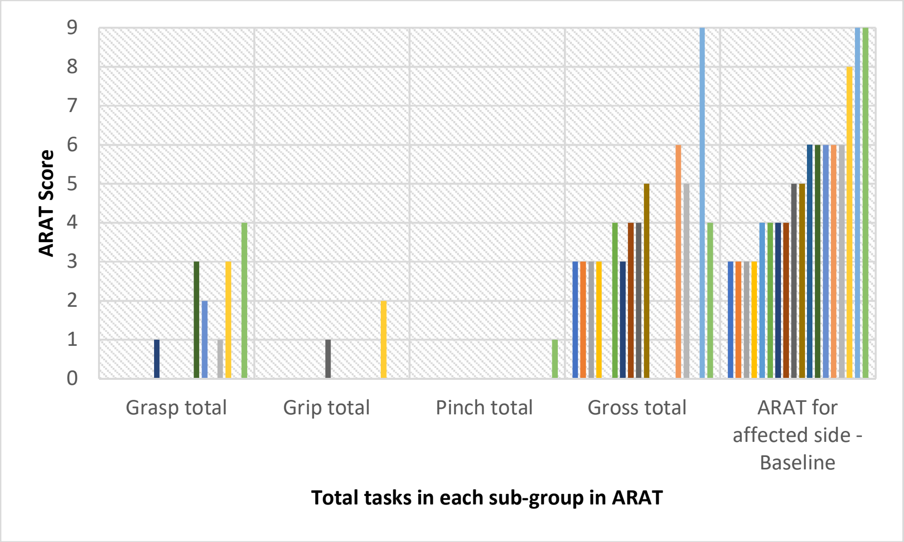


Figure 4-9 Showing individual scores for each task in the ARAT outcome measure. The score of each patient in each sub-group of ARAT is plotted on the x-axis and the total score of ARAT is on the y-axis.

Within the subset, patients who scored between one and nine: Eight (44% with 95% CI 22% to 69%) were able to carry out simulated grasping tasks. Two (11% with 95% CI 2% to36%) could carry out simulated grip tasks. One (1.5% with 95% CI 0.3% to 3%) patient achieved a score in the simulated pinch sub-category.

4.3 Clustering trajectory of ARAT scores:

This section explains the trajectory of the ARAT outcome over three months for the recovery upper limbs. In this part, the K-means method was used to group the scores of ARAT. This method, statistically, needs to determine the number of groups and initial centres. For this reason, the Silhouette method was applied to identify the number of clusters/groups of the ARAT scores. The idea behind using K-means is to produce homogenous clusters/groups that contain similar subjects/ patients. This could help to develop a more accurate prediction model to be used for each group independently.

Cluster analysis was deduced based on the steps as follows: firstly, I applied Silhouette values evaluation method for determining the optimal number of clustering of the ARAT score(Chen et al., 2002; Rousseeuw, 1987). It appears from the Figure 4-10 that the largest average of Silhouette is (0.8) which means that the four clusters have the best number of clustering is four. Selecting four clusters is a good number of clustering to give the accurate result of using partitioning clustering methods. The finding is consistent with the findings of the earlier study by Stinear et al. (2012) which has found the same optimal number of clustering baseline ARAT score. Secondly, after the optimal number of clusters had been identified, K-means was applied to cluster the differences in slops/ rates of change of the ARAT scores respectively, as follows:

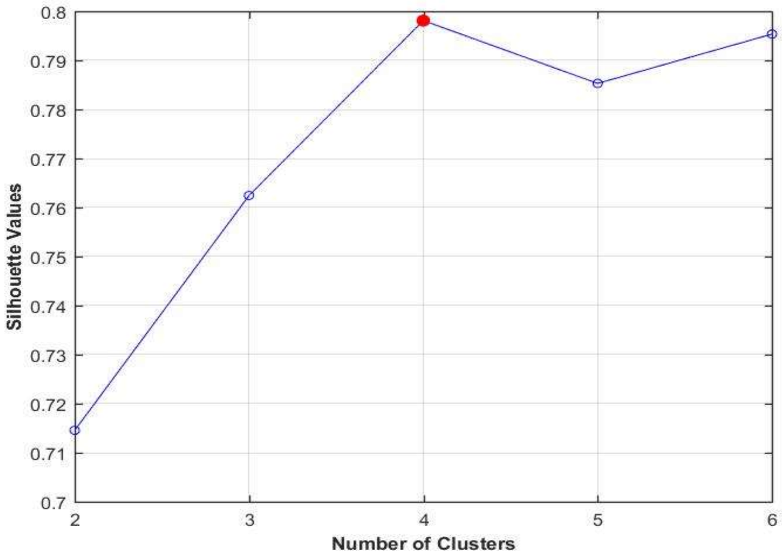


Figure 4-10 Showing the curve elbow of Silhouette method.

a. I found the differences between the baseline line measure of ARAT scores and the four weeks measure, and the differences between four weeks and three months.

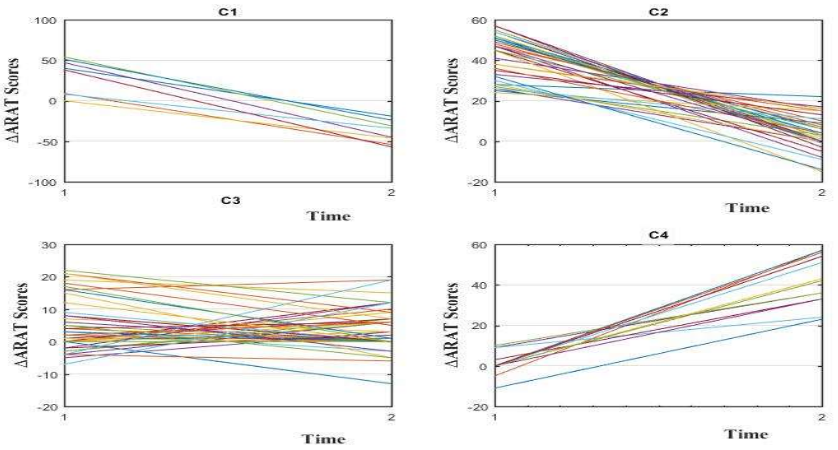


Figure 4-11 Represents the result of K-Means clustering. The trajectory of each patients overtime is plotted on the x-axis and the total score of ARAT is on the y-axis.

b. The next step I depended on the clustering results in step (b) to cluster the three measures of ARAT scores by the same method (K-means), and the analysis of the grouped results show the following:

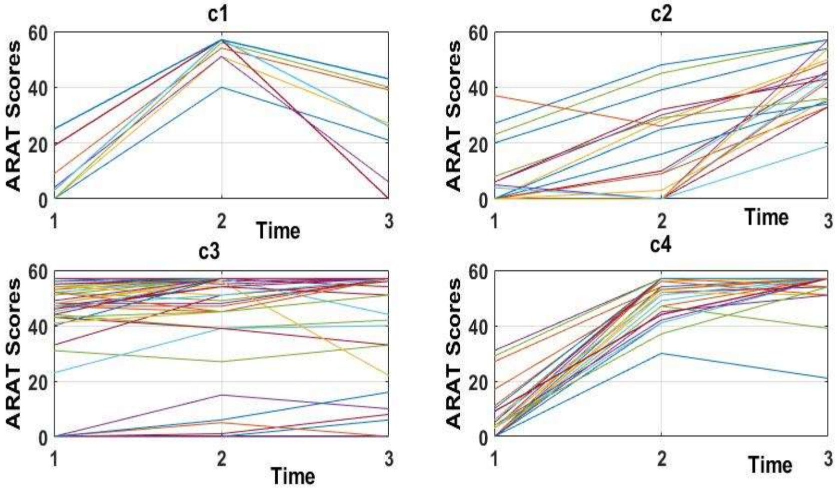


Figure 4-12 Shows the result of K-Means clustering methods on the ARAT outcomes. The trajectory of each patient's overtime is plotted on the x-axis and the total score of ARAT is on the y-axis.

4.4 Discussion

In this chapter, I have described how a cut-off point has been identified to be seven for dichotomising the outcome of the Action Research Arm Test. Firstly, the score seven was selected because some of the patients could partially perform some of the easy tasks such as grasping and one or two tasks of gross movements like patient P11, P15, P16 and P18, as a result in Table 4-1. Previous studies used cut-off point (ARAT>=10) to dichotomize the outcome of ARAT as binary (0,1). One for those who regained some hand and arm and zero for those who did not recover the hand and arm function (Kwakkel et al., 2003; Nijland et al., 2010c). They used a score of 10 as a cut-off point to find the probability of the recovering upper limb after 6 months in people with a flaccid upper limb post-stroke. Additionally, they reported that lower cut-off point of ARAT score might lead to false positives for the return of dexterous precision gripping using the hand and fingers because a low cut-off score only captured the presence of gross shoulder and elbow movements(Kwakkel et al., 2003). Whereas, as results in figures (4-2, 4-3,4-4,4-5,4-6,4-7,4-8 and 4-9) the patient of score nine or less on the ARAT can incompletely perform some of the easy tasks of the grasping part. This reason is mainly reflected in the decision that the cut-off points of the score seven might give a precise dichotomous classification of whether the patients will recover and be independent in their life or not. Moreover, it might help to balance the cost-effectiveness with interventions provided for patients.

The second part of this chapter explains how to show the trajectory of patients' ARAT outcomes from a measurement made three times. The clustering results show the four subgroups of the ARAT. These scores were measured at three-time points over a three-month period (baseline, four weeks and three months). When using K-Means clustering for analysis, patients' scores in each subgroup demonstrated better homogeneity compared with before clustering.

Our aim was to determine which predictor variables will be affected in each cluster and to develop a prediction model of that cluster separately. Due to the limitations of the reduced sample size it was not possible to develop a prediction model of each cluster.

4.5 Conclusion

Patients with scores of less than five could only perform the easiest items within each of ARAT's subgroups other than the gross movements subgroup. For these patients, I can perhaps classify them as having no useful arm function. Patients with a cut-off score of nine can carry out simulated activities that reflect recovery of useful arm function. If a cut-off of nine is used, then there is a risk that patients with recovery potential are missed from receiving treatment. If the cut-off was reduced from nine to seven, then the chances of inappropriately classifying a person as having no useful arm function are reduced. I would therefore conclude that the ARAT cut-off in acute stroke patients should be seven and not nine.

Chapter five

5 Predictors selection

5.1 Introduction

This chapter aims to illustrate the process undertaken to study predictors selection method, using the traditional methods (Univariate variable selection, Stepwise Regression) and penalised methods (LASSO, adaptive LASSO and group of LASSO). However, there are several steps to be followed before discussing the possible predictors selection methods. These steps aim to prepare the dataset before applying methods, hence pre-processing. The pre-processing operations will produce datasets, which requires us to descriptive analyse. The section of processed data will describe and test for multicollinearity. This would be followed by producing a matrix design of the processed datasets. Results of the different predictors selection methods will be presented into the two main categories previously mentioned: classical and penalised methods. Furthermore, I will evaluate the results of each method and compare their performance to identify the best method to be used in this study. Finally, I will discuss the findings of this chapter.

5.2 Pre-processing data

One of the essential aspects of building a model is data pre-processing, which has two steps. The first step is to present one data-set of a retrospective study, with an emphasis on its' variables whether it was the dependent variable (outcome) or independent variables (predictors). The second aim is to test for the level of multicollinearity within the data-set. Pre-processing has several applications; however, only the first, third and fifth applications were used in this study:

- to remove some unsuitable information

- to handle the missing data
- to recode the categorical variable predictors using the dummy variable or designs variables
- to extract some variables to represent the interaction among the predictors
- to test and handle multicollinearity among the predictors

Most prediction models used for prediction of recovery post-stroke are based on a set of data that possibly has high levels of multicollinearity. Consequently, multicollinearity will have a large negative impact on the performance of these models. Based on the literature, a large proportion of clinical outcomes measure the same ICF domains, namely Body structure and function, activity and participation. This could be a contributing factor for multicollinearity between the predictors in the data set. Additionally, the increase in the dataset's dimensions corresponds to the number of patients in the study sample.

5.2.1 Pre-processing of a retrospective data-set of an RCT

A secondary anonymised retrospective data-set was taken from a previous randomised control trial (Church et al., 2006). The RCT's inclusion criteria included participants who had a sustained upper limb problem within the previous/last ten days after acute stroke. Then, the RCT's primary outcome measure, the ARAT, was used on three occasions. In addition, ARAT was collected among other baseline assessments, including other demographic information, hand dominance, the severity of the stroke and stroke subtype (Church et al., 2006). The second and third measures of ARAT were undertaken after 4-weeks and 12 weeks intervention respectively. Additionally, a total of 178 patients and related 249 variable candidates (predictors) were present in the dataset. These 249 predictors comprised of both categorical and continuous data. The RCT's data-set had the issue of missing data and low variability in some predictors.

To handle the predictors with low variability, I must check data on two levels: the individual predictor or the values of a whole record (each patient). The first level delivered information about each predictor in the data-set using some simple statistical tools. For example, find the distribution table of each predictor for checking the predictor's distribution. Then, the predictor was removed if it had less than 90% variability, see Table 5-1.

Table 5-1 Variables removed for being less than 90% variability.

Predictors Name	**Labels**	**Frequency**	**Percent**
Brainstem/cerebellar signs	Yes Left	2	1.1
	Yes Right	3	1.7
	No	171	97.2
	Total	176	100.0
Others deficit	Yes Left	1	.6
	No	175	99.4
	Total	176	100.0
Pre-stroke pain in last month	Yes	8	4.5
	No	168	95.5
	Total	176	100.0
Star cancellation test done	Done	Done	176
NIH Stroke Scale Pupillary response	0	175	99.4
	1	1	.6
	Total	176	100.0

5.2.2 Dealing with missing data

Dealing with missing data in most models is divided into case removal or imputation of a numeric value for the missing value. Case removal is very wasteful of data as it is very rare to find all the information that relates to a patient. This situation usually happens when handling medical data, since it is either not necessary to collect or not important to record a value. This might cause some values in data to be missed. In addition, case removal could change the patient group who might be used for model development and hence the model will be only applicable to the patients who have complete information, see the example in Table 5-2. This means the model is extremely impractical. Therefore, estimating of missing values could be necessary prior to developing the prediction model.

Table 5-2 Predictors have more than 80% missing data.

Predictors Name	Labels	Frequency	Percent
Previous stroke same side affected	yes	13	7.4
	No	19	10.8
	Total	32	18.2
	Missing Data	144	81.8
Pre-stroke pain-Which arm	Right	3	1.7
	Left	2	1.1
	Both	3	1.7
	Total	8	4.5
	Missing System	168	95.5

There are two main statistical methods used to estimate the missing values. These two methods are categorised relative to the value that would replace the missing values. While handling data, if the same value were used to replace all missing data, then a single-imputation method was implemented. Single-imputation methods, depending on the predictors' data type, can be further divided into: 1) For quantitative predictors, the univariate mean value of non-missing values would replace all the missing values of that predictor; 2) For predictors with categorical data, the univariate mode value of non-missing predictors to replace all the missing values of that predictors. Despite the advantages of these processes (mean or mode substitutions) which are simple to apply and quick, one obvious objection would be deflation the variance (variability) that might be undesirable and the bias incorporated into the model by this approach Sterne et al. (2009) .

To overcome this problem of deflation, scholars utilise another univariate imputation method that includes an additional stage. A binary column is added with each predictor containing missing values. The idea behind this is to assess the impact that might occur because clinicians do not usually report the normal value, the imputed values would be biased towards abnormal values. However,

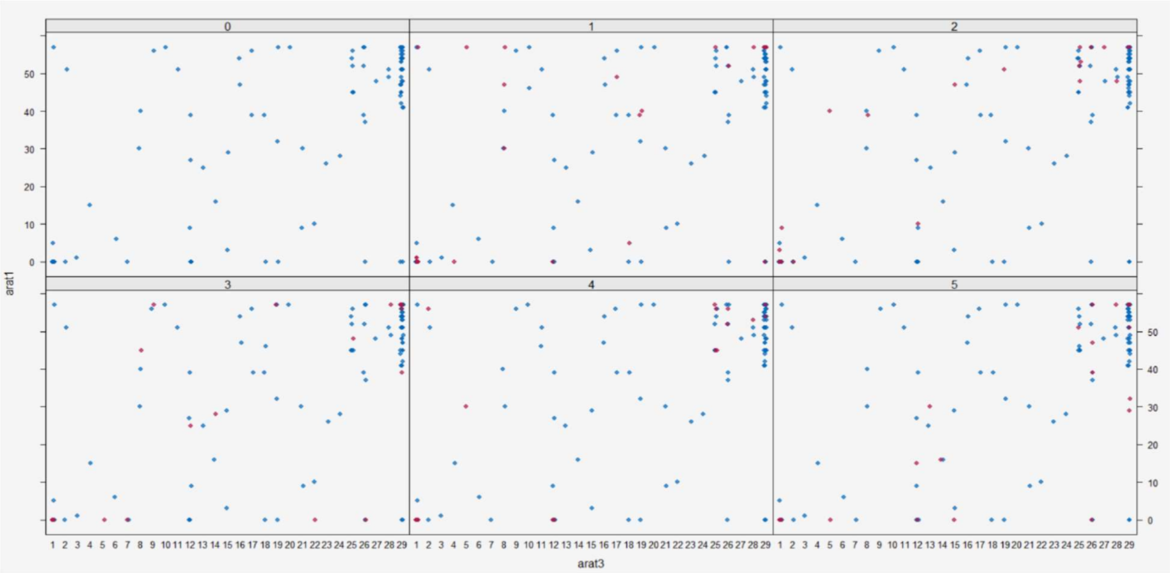


Figure 5-1 Result of imputation missing value using multi-imputation methods

two limitations of this method are: it increases the dimensionality of data, and it would duplicate the number of predictors.

The second category of statistical methods used to estimate the missing values is the multi-imputation methods (MI). MI methods involve repeatedly imputing missing-values and analysing each dataset after every single imputation (Steyerberg, 2009). After each imputation, the average of the overall dataset is computed and used to replace the next missing datum. This will be repeated until all missing data are replaced. The advantage of MI methods over single-imputation methods is that they allow for uncertainty on the missing values by generating numerous imputed datasets and appropriately emerge results from each of them. In our research, the predictive mean matching MI method was used to handle the missing issue (Harrell Jr, 2015).

5.2.3 Outcome's cut-off point

As this pre-processing application has been already discussed in chapter two, this section would contain only this study's implementation of this application. In this study, the outcome ARAT was adopted as response variable (dependent). Since the total of ARAT scores ranged between zero and 57, it is required to transform these scores to the binary values to be able to perform logistic regression models. The dichotomisation was done as follows, a score of zero was given for those who had ARAT scores of less than seven (ARAT <7), and a score of one for those who had ARAT scores greater or equal to seven (ARAT≥7). This means that seven is the cut-off point because a score of ≤ 6 points indicates that the patient can only partially perform the easy task in each subgroup of hand or arm function. Finally, the dataset contains 176 rows (patient) and 76 columns (variables). The next section will describe the produced datasets after pre-processing applications are completed.

5.3 Processed data

5.3.1 Description of processed datasets

The total of predictors includes: 12 demographic variables; nine predictors measures of motor activity; three measures of participation and four predictors representing pain measurements, see Table 5-3.The other group of predictors includes: 25 predictors measures of motor impairments, five predictor measures of cognitive impairments, three predictors measures of visual impairments, eight predictors obtained from the resulted measures of sensory impairments and four predictors measures of speech impairments, shown in Table 5-4.

Table 5-3 Shows the Participant Characteristics and Clinical Measurements of sample size = 176 Patients.

Demographic and historical predictors	*Measures of Motor activity*	*Measures of Participation*	*Measures of Pain*
Age 71.3± 11.3 Gender: Male 78(52%) First stroke 125(83.3%) Days from current stroke to measurement 5.30± 2.59 Days from stroke to admission 0.20± 1.39 Left or right handed 1.15 ± 0.73 Known Diabetes 23(15.3%) Side affected by stroke 95(63.3) Stroke Subtype TACS/PACS vs POCS/LACS 69(46%) Stroke subtype: TACS 44(29%) PACS 37(24.7) LACS 66(44%)	 Sheffield total 17.2±3.6 Motricity Total left leg 75.30± 29.84 Motricity Total right leg 87.10±22.09 Motricity Total left arm 67.41±37 Motricity Total right arm 83.53±27.08 Arm Motricity for affected 51.44±32.69 side -Baseline Leg Motricity for affected 63.53±27.38 side -Baseline Total Motricity for affected 57.60±27.38 side -Baseline Frenchay Arm test for affected side -Baseline 1.77±2.05	Total Barthel score 10.03±5.75 Baseline Barthel (coded) 1.65±1.15 Nottingham EADL Total 15.77±3.70	Pre-Stroke Pain (5) for affected side -Baseline 138(92%) Pre-Stroke Pain (10) for affected side -Baseline 146(97.3%) Post Stroke Pain (5) for affected side -Baseline 115(76.67) Post Stroke Pain (10) for affected side -Baseline 115(76.67)

Table 5-4 Shows the Outcomes clinical measures of (Motors, Cognitive, Visual, Sensory and Speech) impairments.

Measures of Motor Impairments		***Measures of Cognitive Impairments***	***Measures of visual Impairments***	***Measures of Sensory Impairments***	***Measures of Speech Impairments***
Total NIHSS	1.76±1.01				Sheffield-Receptive total
NIH Arm for affected side -Baseline	8(5.33%)	NIHSS of consciousness	Visuospatial deficit	Upper limb Sensory- affected	9.83±2.45
NIH best motor leg-affected side-Baseline	19(12.7 %)	0.15±0.36	at Baseline	side-Baseline	Sheffield -Expressive total
NIH Limb ataxia-affected side- baseline	0.63±0.88		61(40.67%)	66(44%)	7.12±3.62
NIHSS Facial palsy	36(24%)	NIHSS Neglect R	NIHSS Best visual	Sensory deficit affecting	NIHSS Dysarthria
NIHSS Best Motor R-arm	46(30.67)	103(68.67%)	114(76%)	arm/hand	63(42%)
NIHSS Best motor -L-arm	102(68%)	NIHSS Neglect L	NIHSS Best gaze	45(30%)	NIHSS Best language
NIHSS Best motor -R-leg	57(38%)	127(84.67%)	143(95.33%)	Sharp-dull discrimination	125(83.33%)
NIHSS Best motor -L-leg	61(40.67%)			deficit	
NIHSS Plantar reflex	61(40.67%)	NIH Neglect- affected			
NIHSS Limb Ataxia R	38(25.33%)	side- Baseline		64(42.67%)	
NIHSS Limb Ataxia L	143(95.3)	102(68%)		Hot-cold discrimination deficit	
NIHSS Limb Ataxia affected side	137(91.33)	Abbreviated mental test score		54(36%)	
NIHSS grouped	1.39±0.49	Total		NIH Sensory- affected	
NIHSS grouped	1.03±0.18			side-Baseline	
NIHSS grouped	0.65±0.48	61±1.91		70(46.66%)	
NIH Arm for affected side -Baseline	12(8%)				
NIHSS Best Motor L leg	101(67.33%)			Sensory symptoms	
Left shoulder shrug	1.40±0.70			45(30%)	
Right shoulder shrug	1.65±0.58			NIHSS Sensory R	
Shoulder Shrug at Baseline	145(96.6%)			22(12%)	
Passive range of pain-free movement	112(74.67)			NIHSS Sensory L	
The active range on pain-free movement	98(65.33)			48(32%)	
Unilateral weakness affecting arm/hand	1.26±0.53				
Shoulder shrug for affected side -Baseline	125(83.33%)				

5.3.2 Testing multicollinearity

The VIF was calculated to detect the multicollinearity level for assessing the ability to use the traditional methods and LASSO (Dormann et al., 2013). To measure the existence of the multicollinearity problem, the VIF was applied using randomly selected linear combinations, which were among predictive variables within themselves.

Table 5-5 Results of Variance Infraction Factors among the predictors.

	Predictors variable	***Variances Inflation Factors***
1	Days from stroke	1.615
2	Days from stroke to admission	1.582
3	Left or right handed	2.499
4	Previous stroke	1.633
5	Side affected by stroke	26.721*
6	Stroke Subtype TACS/PACS vs POCS/LACS	22.604*
7	Stroke subtype	15.613*
8	Abbreviated mental test score Total	4.041
9	Sheffield total	8.917
10	Motricity Total right arm	9.900
11	Total Motricity for affected side -Baseline	16.393*
12	Frenchay Arm test for affected side-Baseline	28.388*
13	Baseline Barthel (coded)	24.532*
14	Nottingham EADL Total	30.064*
15	Pre-Stroke Pain (10) for the affected side -Baseline	1.598
16	Post-stroke Pain (5) for the affected side -Baseline	17.736*

17	Post-stroke Pain (10) for the affected side -Baseline	17.798*
18	Shoulder shrug for affected side -Baseline	23.140*
19	NIH Stroke Scale Best motor -R-arm	7.419
20	NIH Stroke Scale Best motor -L-leg	15.286*
21	NIH Stroke Scale Plantar reflex	16.546*
22	NIH Stroke Scale Limb Ataxia R	3.587
23	NIH Stroke Scale Limb Ataxia L	2.130
24	Total NIH Stroke score	41.532*
25	NIH Arm for affected side -Baseline	8.290
26	NIH Best motor leg-affected side -Baseline	18.668*
27	Passive range of pain-free movement	5.232
28	Active range of pain-free movement	4.889
29	Left shoulder shrug	1.800
30	Unilateral weakness affecting arm/hand	3.869
31	NIH Best motor leg-affected side -Baseline	25.819*
32	NIH Arm for affected side -Baseline	11.969*
33	Right shoulder shrug	16.502*
34	NIH Stroke Scale Neglect R	6.490
35	NIH Stroke Scale Neglect L	2.801
36	NIH Stroke Scale Level of consciousness	58.775*
37	NIH Stroke Scale Best gaze	2.210
38	NIH Stroke Scale Best visual	4.285
39	NIH Stroke Scale Sensory L	20.697*
40	Visuospatial deficit at Baseline	8.129
41	Sensory deficit affecting arm/hand	254.278*

42	Upper limb Sensory deficit at Baseline	4.573
43	Hot-cold discrimination deficit	3.536
44	Sharp-dull discrimination deficit	19.409*
45	NIH Sensory-affected side -Baseline	190.769*
46	Sensory symptoms	75.868
47	Sheffield -Receptive total	5.482

* *high level of Multicollinearity*.

The results of the VIF test show that there was evidence that the data set had a high level of multicollinearity. The decision was made based on the rule of VIF >10 that means a high level of multicollinearity existing among predictors.

5.3.3 Matrix design

This section describes the method used to build two design matrices used for model training and model testing. Both design matrices will have the same columns (predictors), but the number of rows (patient observations) in each matrix is not necessarily equal in both matrices. Data was split randomly into 70% and 30% for the training and testing phases, respectively. Training phase results will represent the selection predictors process; whereas, testing phase results will represent evaluation of the performance of the models' selections.

5.4 Process of Predictors Selection and Methods

In this section, the results of selection predictors process are presented based on the method used to select them: stepwise logistic regression method, and penalised methods. Before applying the stepwise logistic regression method, two steps were performed: univariate and multivariate logistic regression methods. Univariate logistic regression method was used for determining the relationship between each predictor at a baseline level and outcomes variable after three months. For categorical predictors, the contingency table test of ARAT outcomes (0,1) versus the (k) levels were used to find the likelihood ration Chi-square test with (k-1) degree of freedom. In the case of continuous predictors, the univariate logistic regression analysis was applied to fit the model and calculate the likelihood ratio test and the Wald test. The predictor was eliminated where it had a non-significant association at level (α=0.05). After all predictor candidates were identified for the inclusion of univariate of logistic regression, multivariate logistic regression was fitted. Finally, the stepwise logistic regression method was used to eliminate all predictors variables of the model that have not had any contribution to explain the effect of the outcome. All the previously mentioned steps represent the models' training set.

The penalised methods of predictor selection include three methods: adaptive LASSO, Group LASSO and LASSO. During applying penalised methods, the tuning parameter need to be estimated. The cross-validation approach was used to account for the tuning parameters of the penalty value in ALASSO, GLASSO and LASSO. A Tuning parameter acts as a regulator between amplitudes of bias versus prediction error (or variance) in the model. Cross-Validation was applied by dividing training data into equal k-folds, which are randomly-selected sub-samples. k-1 folds of these were allocated to the model's training phases, and the remaining fold was used in the model's testing phase. ALSSO, GLASSO and LASSO were then applied to the training data for a range of different values of estimated tuning parameter. Each fitted model was then used to predict the outcome of the ARAT in the test fold, recording the prediction's deviance (mean square error) for each value of the tuning parameter. This process is repeated iteratively, resulting in all ten parts of the data being used for estimating the penalty value, which is used to shrink the candidate predictors' variables and choose the essential subset of predictors. In order to estimate the confidence intervals for the coefficients obtained from the modelling process, Bootstrapping was used. In brief, the confidence intervals for each variable were estimated from resampling the data to produce 500 data sets each comprising of 176 patients (reflecting the size of the original sample size) using an empirical distribution function. Penalised methods were then applied to these 500 data sets and from this the 95% confidence intervals were estimated. The confidence intervals of the estimated coefficient calculated using the exponential values of 2.5th percentile, mean values and 97.5th percentile for each coefficient estimated using penalised methods(Efron and Tibshirani, 1994).

Finally, testing phase aimed to compare the performance of four models of data set. The calibration was performed by plotting the calibration curve of four of the developed models. The discrimination was assessed by using the C statistic. The confusion matrix was used to evaluate the fit of the four models to see what rate of true positives is classified as being positive (the sensitivity) and what rate of true negatives is classified as being negative (the specificity). The discriminative ability of the four models for the upper limb recovery was calculated by measuring the area under the ROC curve and plotted.

5.4.1 Classical methods

The classical methods' results, which were introduced in section 3.2, presented in this section are: Univariate and stepwise logistic regression selection methods.

5.4.1.1 Univariate logistic regression selection results

All baseline predictors were assessed in a univariate analysis. Based on the univariate analysis in Table 5-6 , 51 predictors at baseline were associated significantly with Action Research Arm Test that measure after three months (all P<0.05), for example, the older patients had a lower chance of upper limb recovery than the younger (p< 0.05).

Table 5-6 Results of the univariate logistic regression models.

No.	***Predictors***	***Estimated coefficients***	***Variance***	***P-Value< 0.0001***	***Null deviance***	***Residual deviance***	***AIC***
1.	Age	-0.068	0.020	0.00005	197.94	195.74	199.74
2.	Unilateral weakness affecting arm/hand	1.6094	0.5560	0.0001	197.94	175.82	179.82
3.	Sheffield total score	0.12234	0.03862	0.001	197.94	192.78	196.78
4.	Motricity Total left leg	0.07711	0.006438	0.0005	197.94	187.67	191.6
5.	Motricity Total left arm	0.038177	0.005788	0.0005	197.94	196.74	200.7
6.	Motricity Total left side	0.045149	0.007007	0.0001	197.94	196.73	200.7
7.	Left shoulder shrug	1.6614	0.2971	0.0001	197.94	197.33	201
8.	Active range on pain-free movement	0.043340	0.008865	0.0001	197.94	197.46	201.46

No.	*Predictors*	*Estimated coefficients*	*Variance*	*P-Value< 0.0001*	*Null deviance*	*Residual deviance*	*AIC*
9.	Hot-cold discrimination deficit	-1.1299	0.3591	0.0001	197.94	145.24	149.2
10.	Sharp-dull discrimination deficit	-0.9904	0.3605	0.006	197.94	187.79	191.79
11.	Total Barthel score	0.38986	0.06359	0.000	197.94	197.94	201.94
12.	NIH Stroke Scale Level of consciousness - questions	-0.9129	0.3379	0.007	197.94	120.81	124.81
13.	NIH Stroke Scale Best visual	-0.8983	0.2004	0.000	197.94	195.15	199.15
14.	NIH Stroke Scale Facial palsy	-0.7903	0.1743	0.000	197.94	177.41	181.41
15.	NIH Stroke Scale Best motor -R-leg	-1.0888	0.1937	0.00005	197.94	196.43	200.43
16.	NIH Stroke Scale Best motor -L-leg	-1.0780	0.3309	0.001	197.94	159.10	163.1
17.	NIH Stroke Scale Sensory R	-1.0928	0.2782	0.0001	197.94	193.93	197.93
18.	NIH Stroke Scale Neglect L	-1.0805	0.2990	0.0001	197.94	167.44	171.44
19.	Total NIH Stroke score	1.6392	0.7561	0.03	197.94	127.31	131.31

No.	*Predictors*	*Estimated coefficients*	*Variance*	*P-Value< 0.0001*	*Null deviance*	*Residual deviance*	*AIC*
20.	Baseline Barthel (coded)	0.050395	0.007764	0.00005	197.94	193.82	197.82
21.	Arm Motricity for affected side at Baseline	0.057417	0.008963	0.0000	197.94	133.08	137.08
22.	Total Motricity for affected side -Baseline	0.8360	0.1945	0.0000	197.94	135.14	139.14
23.	Sensory symptoms	-1.1481	0.2226	0.0000	197.94	190.16	194.16
24.	NIH Limb ataxia-affected side -Baseline	-0.6971	0.2616	0.007	197.94	194.57	198.57
25.	NIH Sensory-affected side -Baseline	-1.0262	0.2051	0.000	197.94	190.76	194.76
26.	NIH Score grouped	2.3286	0.5065	0.000	197.94	196.78	200.78
27.	NIH Score grouped	-2.3906	0.5068	0.000	197.94	167.10	171.1
28.	NIH Score grouped	2.26383	0.420	0.000	197.94	165.24	169.24
29.	NIH Score grouped	-2.2638	0.4203	0.000	197.94	162.14	166.14
30.	NIH Score grouped	-2.0794	0.3992	0.000	197.94	162.14	166.14

No.	***Predictors***	***Estimated coefficients***	***Variance***	***P-Value< 0.0001***	***Null deviance***	***Residual deviance***	***AIC***
31.	Shoulder Shrug at Baseline	-1.8412	0.5062	0.0003	197.94	175.42	179.42
32.	Upper limb Sensory deficit at Baseline	-1.8606	0.3951	0.000	197.94	193.91	197.91
33.	Visuospatial deficit at Baseline	-1.2541	0.3997	0.0017	197.94	172.42	176.42
34.	NIH Neglect-affected side -Baseline	14.4827	1029.121 5	0.989	197.94	170.78	174.78
35.	NIH Best motor leg-affected side -Baseline	1.5060	1.0338	0.145	197.94	165.92	169.92
36.	Pre -Stroke Pain (10) for affected side -Baseline	13.4751	882.7434	0.988	197.94	197.52	201.52
37.	NIH Arm for affected side -Baseline	-0.07819	0.11652	0.502	197.94	138.90	142.9
38.	Baseline Barthel (coded)	-0.1118	0.3839	0.771	197.94	143.76	147.76
39.	Baseline Barthel (coded)	17.7964	1072.315 2	0.987	197.94	197.86	201.86
40.	Baseline Barthel (coded)	15.5219	906.9427	0.986	197.94	173.54	177.54
41.	NIH Stroke Scale Dysarthria	-0.1601	0.2761	0.562	197.94	183.48	187.48

No.	***Predictors***	***Estimated coefficients***	***Variance***	***P-Value< 0.0001***	***Null deviance***	***Residual deviance***	***AIC***
42.	NIH Stroke Scale Best motor -R-arm	0.2468	0.1967	0.21	197.94	172.96	176.96
43.	NIH Stroke Scale Best motor -L-arm	0.2732	0.2311	0.237	197.94	148.69	152.69
44.	Motricity Total right side	-0.008856	0.007905	0.262	197.94	139.04	143.04
45.	Sensory deficit affecting arm/hand	1.020	1.075	0.343	197.94	181.47	185.47
46.	Side affected by stroke	-0.09531	0.61170	0.876	197.94	172.45	176.45
47.	Post-stroke pain -10-point scale	0.08841	0.60027	0.883	197.94	193.50	197.5
48.	Motricity Total right arm	-0.007246	0.006845	0.289	197.94	139.28	143.28
49.	Motricity Total right leg	-0.008805	0.008379	0.293	197.94	154.65	158.65
50.	Passive range of pain-free movement	-0.004650	0.006894	0.500	197.94	162.38	166.38
51.	NIH Stroke Scale Plantar reflex	1.0257	0.5972	0.08	197.94	181.73	185.73

5.4.1.2 Stepwise regression method result

Once the predictors from the univariate logistic regression model was extracted, stepwise of multivariate logistic regression was deduced. The stepwise selection method identified 16 significant predictors at (p-value= 0.05), including the results of selection and fitting as follow:

Table 5-7 Presents the results of the stepwise logistic regression selection predictors based on the AIC.

No.	Name of predictors	Coefficient estimated	Error of estimated	Wald's Test	p-value
1	Age	-0.36	0.13	-2.3	0.02*
2	Total of Sheffield Expressive	-1.59	0.88	-1.92	0.55
3	Total of Sheffield score test	0.64	0.33	1.93	0.05
4	Motricity Total left arm	0.12	0.086	1.424	0.15
5	Right shoulder shrug	9.75	4.62	2.109	0.03*
6	NIH Stroke Scale Level of consciousness	10.301	4.78	2.153	0.03*
7	NIHSS Facial palsy	4.23	1.846	2.293	0.02*
8	NIHSS of best motor- left Arm	2.38	1.459	1.633	0.10
9	NIHSS of best motor- right leg	8.36	.338	2.505	0.01*
10	NIHSS sensory right	9.19	.8558	2.386	0.01 *
11	NIHSS neglected	-16.27	6.928	-2.352	0.01 *
12	Total of NIHSS score	-5.49	2.168	-2.536	0.011 *
13	Baseline Barthel	1.74	1.888	-0.923	0.35
14	Arm Motricity of affected side at baseline time	-0.083	0.0718	-1.161	0.24
15	NIHSS neglect of affected side at baseline	20.93	8.588	2.438	0.0148 *
16	NIHSS score grouped	-9.68	4.184	-2.314	0.02 *

The table above shows the results of predictors selection using the stepwise selection of logistic regression.

5.4.2 Penalised selection results

In this section, estimating the value of tuning (penalty) parameter process and the results of penalised methods selection will be presented individually according to the method used:

5.4.2.1 Estimating tuning parameter

The estimated tuning parameters are eight folds cross-validation, which was applied to estimate the optimal value of penalty (λ) on the training data set. This process aims to obtain held-out performance estimates for the model across a set of possible tuning parameters. The best-achieved value is then selected as the optimal value of the estimated tuning parameters set. This best-identified value will be utilized in the final model that corresponds to the minimum value of deviance (mean square errors).

The flowchart shows the cross-validation approach used to determine the performance of a single value of the set. This process is repeated for all rows of the vector λ, i.e. all possible values of the vector in the grid search.

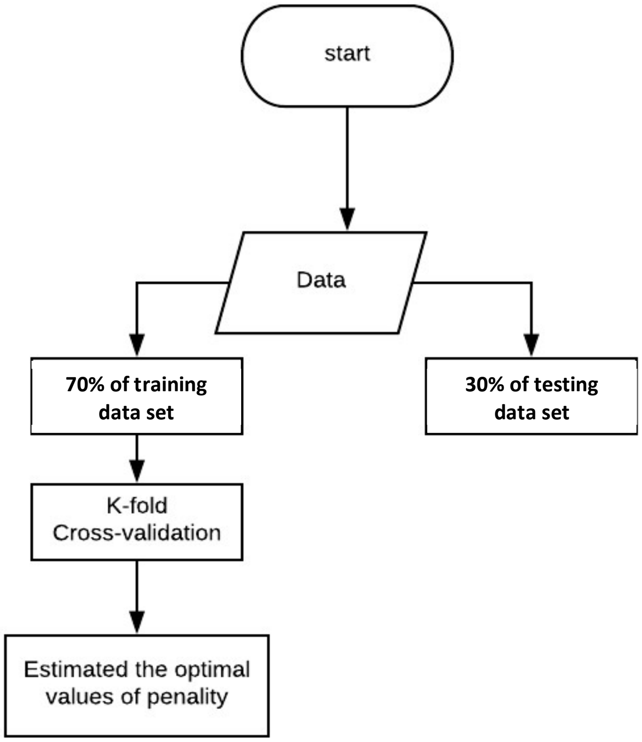


Figure 5-2 Flowchart of estimating process penalty and developing for penalisation methods.

The idea behind the flowchart is to acquire the generalisation performance of all possible model combinations which are: 1) data split to training and testing sets, and 2) training set divided into eight parts/folds cross-validation for model development. This procedure produces 16 values achieving estimates for two values of penalty. These performances are averaged to obtain the overall performance for the optimal estimation value, and the value with the minimum deviance (mean square error) is used for the final predictor's selection and model development.

Data for prediction recovery model were extracted based on the clustering k-means of three times measurement of Action Research Arm Test results. This resulted in a group of 114 patients that included the patients with moderate to severe dysfunction of the upper limb. 70% of the data was taken as a training set to build the model. 30% of the data was used for achieving of the final model. Data were pre-processed as described in the section (5.2) resulting in a design matrix X for training and testing. Three of final models were built to select predictors and prediction based on the RCT data.

5.4.2.2 Adaptive LASSO selection results

The predictors selection process of Adaptive LASSO contains two steps. The first step is to estimate the weight of adaption using the ridge regression method. Then, the value of tuning parameter was estimated based on ridge regression method. The estimated value corresponds to the minimum value of binomial deviance and then the final selection of predictors. The optimal selection is located between the first and the second dash line as in

Figure 5-3 that is corresponding to the minimum value of deviance(Hastie et al., 2015). The optimal selection shrinks unrelated predictors from (74) to 8 predictors by estimating optimal value of tuning parameters (λ =-1.2) of adaptive LASSO. The dotted line on the left side corresponds to λmin. The second line is λ1se.

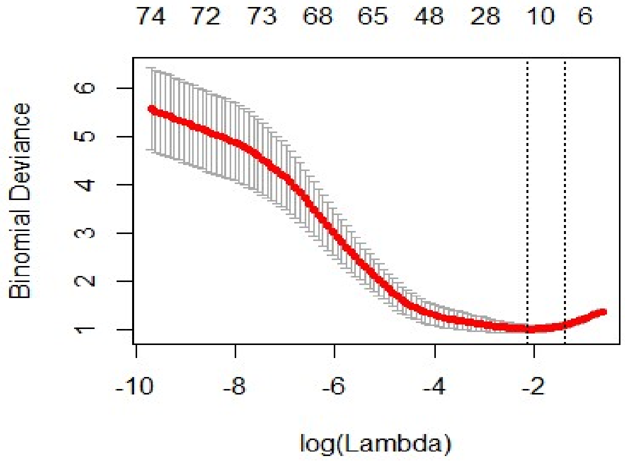


Figure 5-3 Plot of the deviance cross-validation as a function of the penalty parameter λ to determine the.

The results of selection of the adaptive LASSO method include six of related predictors from 74 predictors in the matrix design. The estimated regression coefficients and the confidence intervals of those related predictors are presented in Table 5-8.

Table 5-8 Presents the important predictors selected using the adaptive LASSO variable selection model and the corresponding estimation of regression coefficients; the odd ratio of intercept was 3.58.

No.	**Name of predictors**	**Odd ratio of Coefficient estimated and (estimated 95% CIs)**
1	Motricity Total left the leg	0.0025 (0.0 to 0.005)
2	Motricity Total left the arm	0.0043 (0.0 to 0.005)
3	Total of NIHSS score	-0.56 (-0.105 to 0.00)
4	Baseline Barthel	0.64 (0.00 to 0.749)
5	Arm Motricity of the affected side at baseline time	0.001 (0.00, 0.001)
6	Total Motricity for affected side-Baseline	0.53 (0.0, 2.54)

5.4.2.3 Group LASSO selection results

In the beginning, the optimal value of the tuning parameter was estimated (λ=-2.8) using the 10-fold cross-validation method. This method determines the optimal solution for the number of groups that shrink unrelated groups of predictors from (63) to 8 predictors/groups. The value is located at the dashed line as shown in Figure 5-4. The results selection of the group LASSO method obtained corresponds to the optimal estimated value of the penalty, with the selection method being eight predictors as they are included in Table 5-9.

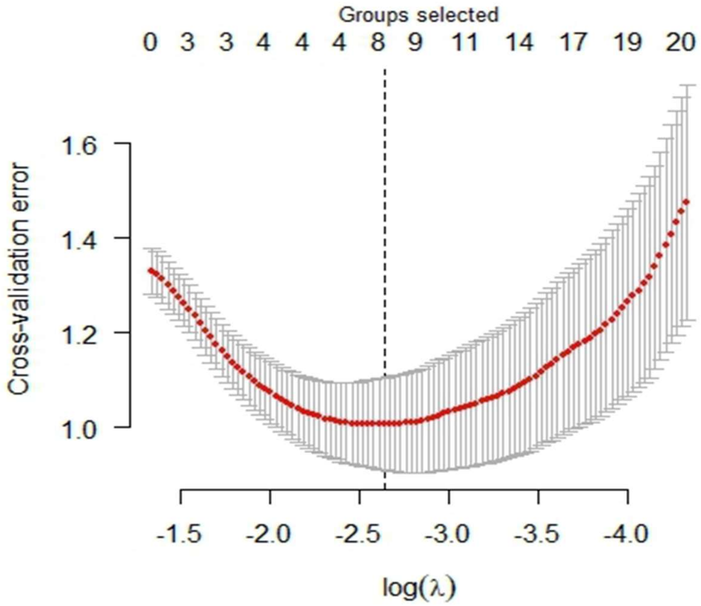


Figure 5-4 Plot of the deviance cross-validation as function of the penalty parameter λ to determine the estimated optimal value of tuning parameters estimated of Group LASSO. The dotted line on the left side corresponds to λmin.

Table 5-9 Shows the significant predictors selected using the Group Lasso variable selection model and the corresponding estimation of regression coefficients.

No.	**Name of predictors**	**Coefficient estimated**
1	Age	-0.03 (-0.22,0.0)
2	Motricity total left arm	0.0148 (0.0075,0.016)
3	Active range of pain-free movement	0.004 (0.00,0.013)
4	Total of Barthel Index	0.15 (0.0,0.12) *
5	NIHSS of the level of consciousness	-0.526 (-0.003,0.00) *
6	NIHSS of neglected left	-0.019(-0.0014,0.00) *
7	Total of NIHSS score	-0.041 (-0.0012,0.00) *
8	Total Motricity for affected side- Baseline	0.0003 (0.0011 to 0.011) *

5.4.2.4 LASSO selection results

In this method, the tuning parameter was estimated to determine the optimal value of the penalty that achieves the best selection of predictors; which reflects the lowest attitude of deviance.

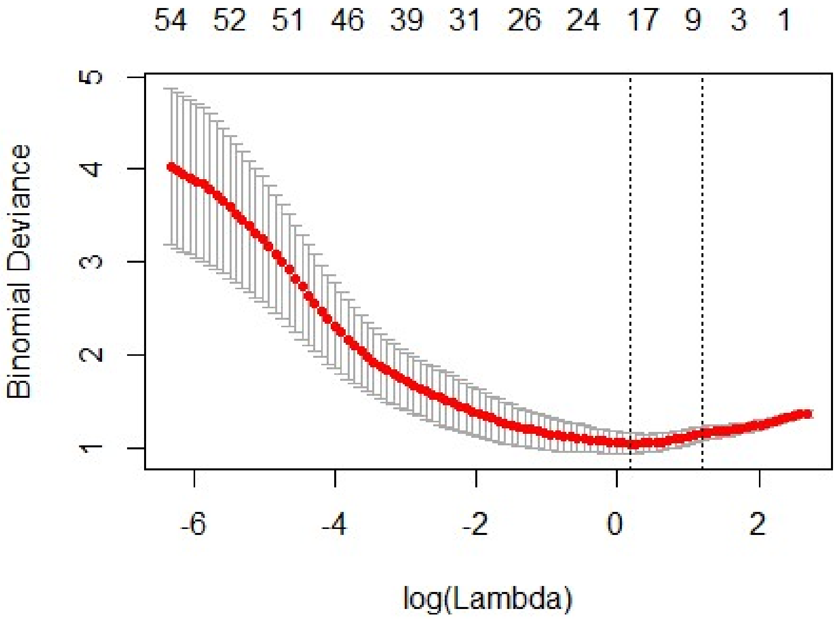


Figure 5-5 Plot of the deviance cross-validation as a function of the penalty parameter λ to determine the estimated optimal value of tuning parameters estimated of LASSO. The dotted line on the left side corresponds to λmin. The second line is λ1se

The optimal tuning parameter λ = 0.0017, corresponding to the minimal deviance of 0.2301, was chosen Figure 5-5. Eight significant variables were estimated from the (74) coefficient paths for the fitted LASSO model based on the optimal (λ). These are listed in Table 5-10.

Table 5-10 Shows the important predictors selected using the Lasso variable selection model and the corresponding estimation of regression coefficients.

No.	Name of predictors	Coefficient estimated
1	Age	-0.024 (-0.57, 0.004)
2	Days from stroke to admission	0.021 (0.0, 1.52)
3	Previous stroke	0.244 (0.00, 2.5)
4	Active range of pain-free movement	0.007(-1.055,0.0) *
5	Total of Barthel Index	0.58 (0.0,0.25) *
6	NIHSS of the level of consciousness	-0.06 (-0.31,0.0)
7	Barthel Index of (0-4)	-0.345 (0.00,0.001) *
8	Total of NIHSS score	-0.048 (-0.31,0.0)

5.5 Evaluation of predictors selection methods

Several numbers of performance measures exist for predicting models. I used measures that are the most common in medical studies of prediction in a medical journal(Steyerberg and Vergouwe, 2014). These measures include the C statistic (The area under the curve of ROC) for discrimination and Brier Score, Hosmer Lemeshow and good-of-fit test for calibration. The results of discrimination and calibration of four models of selection are the next two subsections:

5.5.1 ROC and area under the ROC:

A ROC curve was plotted to determine the discriminatory power of the four models of selection to differentiate between recovered and not recovered, Figure 5-6. As mentioned previously in section 3.3.3, a value of 0.5 means that the model is useless for discrimination (equivalent to tossing a coin) and values near one means that higher probabilities will be assigned to cases with the outcome of interest compared to cases without the outcome.

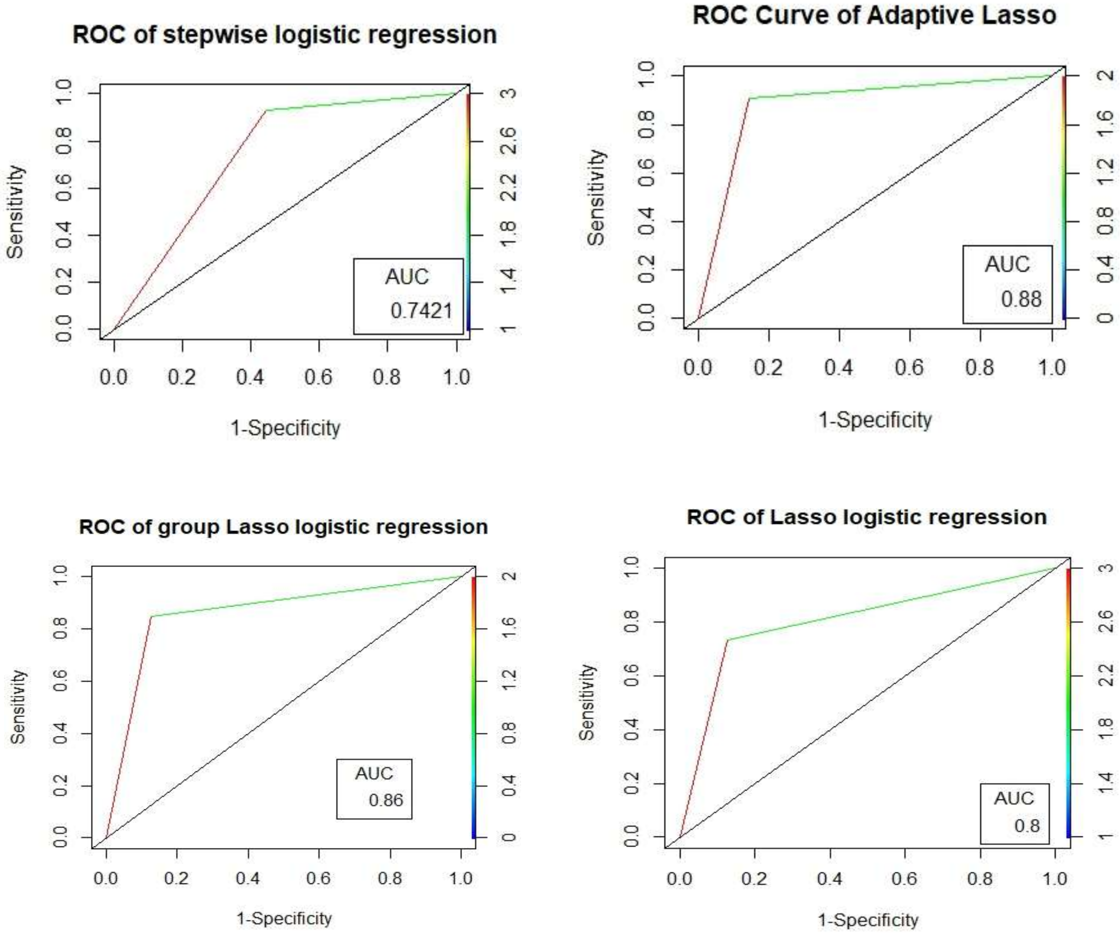


Figure 5-6 Receiver operating characteristic (ROC) curves for the predicted probabilities recovery of the upper limb using four methods (stepwise, adaptive LASSO, group LASSO and LASSO).

The value of the area under the ROC represents the ability of the model to discriminate between those patients who experience a higher probability of upper limb recovery based on the ARAT and those who do not experience the recovery. The area under the ROC was calculated to assess the discrimination for each model of the selection. This was deduced using the pROC and AUC packages in the software R (Kundu et al., 2011; Robin et al., 2011).

Table 5-11 Area under the ROC of four internal validation of model selection.

Methods	**Stepwise model**	**Adaptive LASSO mo**	**Group LASSO**	**LASSO**
Area under the ROC	0.74	0.88	0.86	0.80

The area under the ROC curve for a predicting model is typically between 0.6 and 0.85 (Royston et al., 2009). As I can see, the adaptive LASSO has a higher value of the area under the curve (0.88) than the other methods selected as seen in Table 5-11. This indicates that the ALASSO produces results with good balance between the true positive rate (patient has a probability to recover function of upper limb and, patients who have a chance to recover the function of the upper limb after three months),and the false positive rate (patient, in reality, who has not had a chance to recover but the model identified him/her as the opposite). Furthermore, and since the area under the ROC of GLASSO is (0.86), this means that the GLASSO method can discriminate patients better than the LASSO and stepwise logistic regression.

Additionally, the prediction models with fewer predictors that were identified by the methods of adaptive and group LASSO had the minimum average of prediction errors and the maximal areas under the curve of ROC. This demonstrates that these two methods out-performed the other methods with respect to the identification of the most informative factors.

5.5.2 Calibration

Assessing calibration of four prediction of development models was deduced using two methods; Hosmer Lemeshow test and calibration plot. The Hosmer Lemeshow test was used to investigate how well the predicted probabilities agree with the observed probabilities (calibration). In this test, the way of assessing the fit of a logistic regression model is to compare the expected and observed numbers of positives for different subgroups of the data. This test should not be statistically significant, a p-value is greater than 0.05 showing that the model fits the data. Table 5-12 presents the results of four models.

Table 5-12 Results of Hosmer-Lemeshow test of four methods selections.

Methods	**Hosmer- Lemeshow test**	**P-value**
Stepwise	3.77	0.8
Adaptive LASSO	5.31	0.72
GLASSO	8.718	0.367
LASSO	10.43	0.21

If the result of Hosmer Lemeshow test is non-significant, this means the observed and expected numbers are sufficiently close, then I can assume that I have an adequate model (Steyerberg et al., 2010b). Calibration curves were plotted for each method of selection (Stepwise, adaptive LASSO, GLASSO and LASSO; see Figure 5-7). These curves are performed to plot and show the predicted proportion versus the actual proportion of recovery of function of upper limb of stroke patients based on the ARAT score. If the model was ideally calibrated at each patient, the predicted (dash line) and observed (bold line) values would sit perfectly on each other. In the case of each model for predicting recovery, the predicted and observed lines are not perfectly matched although they are close, especially for ALASSO method. Therefore, there is no reason to doubt that the internal validity of the ALASSO is better than the internal validity of GLASSO, LASSO and Stepwise.

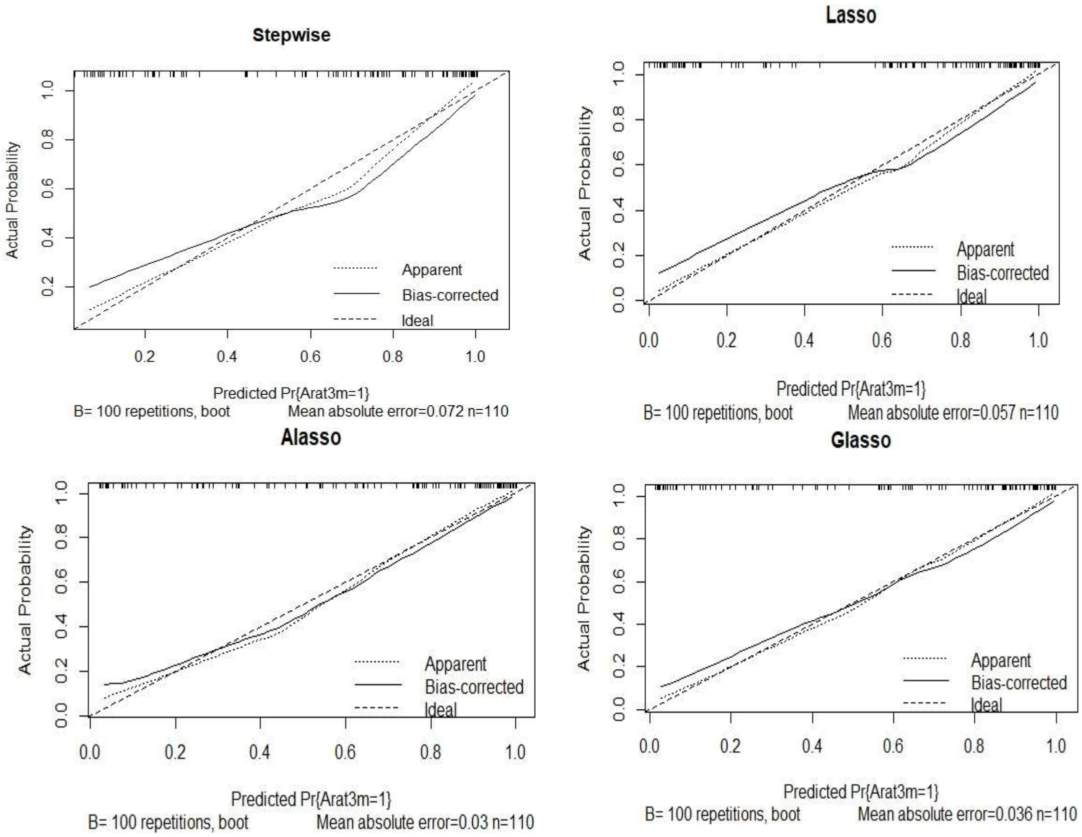


Figure 5-7 Calibration of the predicted probabilities recovery of the upper limb using four methods (stepwise, adaptive Lasso, group Lasso and Lasso).

5.6 Comparison of performance of predictors selection methods

The comparison study presents the comparison of the methods' performance to select relative predictors for developing the model (Table 5-13). Sensitivity and specificity analysis were reported. Sensitivity is the proportion of the true positive outcomes (for example, truly recovered subjects) that are predicted to be positive. Specificity is the proportion of the true negative outcomes (for example, truly not recovered subjects) that are predicted to be negative. A higher sensitivity level indicates better performance of identification of informative predictors from a pool of selected variables.

Table 5-13 Predictive performance of predictors identified by the four methods of selection to distinguish patients which have a chance to recover of upper limb function virus which is not based on the action research arm test after three months.

Penalised methods	Lasso	ALASSO	GLASSO	Stepwise Regression
Sensitivity	1.00	0.91	0.92	0.74
Specificity	0.67	0.86	0.78	0.56
Accuracy	0.86	0.89	0.86	0.60
P-Value	0.006	0.00002	0.0054	0.001
95% CI	(0.664, 0.97)	(0.817, 0.9273)	(0.64,0.97)	(0.1-1.12)

5.7 Discussion

In this chapter predictors selection are applied to RCTs dataset, to build a predictive model and identify the predictors that have an important effect on explaining recovery in the post-stroke upper limb function based on ARAT. Four methods of predictors selection models are investigated when building a model involving stepwise logistic regression and three of penalised methods (LASSO, GLASSO and ALASSO). The best performing selection models, in term of general performance and calibration and the discrimination, was found when the predictors were carried out using the cross-validation and bootstraps methods, choosing the final model by maximising the accuracy of classification. The performance of the models selected compares favourably with the predictive models of post-stroke recovery of upper limb function reported in recent literature.

Theoretical studies and simulations reported that the traditional method of sub-set selection performs poorly whether in the multivariate or multivariable regression model, particularly when (Dormann et al., 2013; Guo et al., 2015; Kwah and Herbert, 2016; Zhang et al., 2015; Zhang et al., 2010):

- Most predicting variables have good explanatory power for the outcome of interest
- Interpretation is complex
- Multicollinearity is present
- The number of predictors is large

The results of this study are supported by this position; thus, traditional methods are not accurate enough methods for predictors selection. In contrast, the penalised methods of selection are reported to have superior performance (Hastie et al., 2015). There are three key findings:

(1) ALASSO method has a better performance when dealing with selecting predictors of upper limbs functional recovery in stroke patients presenting with upper limb motor impairment.

(2) This method has impressive achievement in selection predictors that are clinically relevant to the function of upper limbs' recovery. Selection predictors

based on this method identified predictors which are clinically relevant (NIHSS, Motricity Index, and Barthel Index), which are, according to previous literature, a very good clinical predictor(Coupar et al., 2012; Kwakkel et al., 2003). In addition, NIHSS was found to have broad predictive utility of mortality, disability and independence of ADL. Furthermore, it has been routinely used in acute stroke units.

(3) ALASSO method selected a group by reducing the candidates set of predictors to the suitable subset of the model for predicting the recovery of function of upper limb after 12 weeks, Table 5-8. This will be discussed in the following paragraph. ALASSO selected six predictors and the GLASSO method selected eight predictors. In both the selection methods' results, several predictors/variables are shared, such as the motricity index. This indicates that the motricity index test is a good predictor of the functional upper limb in recovery (Coupar et al., 2012). The MI has shown to high reliability for muscle strength measurement post-stroke. Additionally, it is considered as a simple and easily applied tool that does not need any equipment, training or experience. Thus, it is widely used due to its simplicity and being time and cost-effective. Relative to outcomes, some predictors seem to have a significant explanatory effect; however, they have not been selected. The Barthel index (BI) is a widely used measure for active daily living that contains items covering the most common activities needed for independent living. Moreover, it has shown high sensitivity and specificity in neuroglial conditions in general, and in stroke specifically. (Ohura et al., 2017).

This limitation in the study could be related to the small sample size. This includes the number of missing-data of each patient and the characteristics of each predictor. However, where possible, I wanted to add as many relevant predictors as possible and estimate the missing data with the final data set. Besides this, penalised estimation is a procedure that reduces the variance of estimators by introducing substantial bias. For this reason, I cannot use the same approach in LASSO as in classical logistic of interpreting the predictor's effect estimation.

To ensure the ease of interpretation, a few techniques have been proposed, for example the bootstrap method to estimate the mean square error of estimation and confidence interval. When comparing the results of selected predictors between the ALASSO method and the other methods of selection, Table 5-13, all estimated coefficients of ALASSO are symmetric in the 95% confidence interval. However, some of the estimated coefficients of the GLASSO were located outside of the 95% confidence, indicated by Table 5-9 , Table 5-10 and Table 5-8. This may be due to several reasons, the most likely of which relates to the small sample size of the original sample or some shortcomings of the methods of estimation, or the number the size of each bootstrap sample.

Significant multicollinearity exists within this data set as indicated by the value of VIF which is greater than 10 for many variables(Dormann et al., 2013). Multicollinearity within this data set has arisen from multiple measures being used to assess similar constructs (for example the Frenchay Arm Test and the Motricity Index) and the interdependency between measures of impairment and activity. However, it is also possible that multicollinearity could have been incidental to the inherent nature of the sample (i.e. recovery from stroke follows a small range of predetermined, although as yet undescribed, trajectories), and the small sample size when compared to a large number of independent variables identified. The challenge of using traditional methods in the presence of multicollinearity(Dormann et al., 2013;

Steyerberg, 2009), particularly in small sample data sets, is the risk of selecting inappropriate or confounded predictor variables (for example a latent variable). This problem is resolved within ALSSO and GLASSO methods.

The present study could have a methodological advantage over prior studies, in which I have applied two different types of penalised selection methods (modern methods). As a result, shrinkage of the redundant and irrelevant predictors was achieved in the modelling process for predicting recovery of upper limb function. Use of these methods can also significantly increase the accuracy of the prediction model and allow for easy interpretation of the model. The adaptive LASSO has a better precision score when compared with the LASSO and GLASSO method as per table 4. Both penalised methods were able to select subset predictors in order to predict the outcome of patient's functional upper limb recovery three months post-stroke. Three penalised methods were able to achieve a prediction accuracy higher than 86% sensitivity greater than 93% and specificity between 62% -72%. However, the Stepwise method achieved prediction accuracy 60%, 74% sensitivity and 56% specificity.

5.8 Conclusion

Penalised methods for selection of predictor subsets in prospective modelling of upper limb function, three months post-stroke, have been identified as appropriate. These methods can overcome previous limitations associated with traditional methods, such an existence of a significant correlation between predictors. So far, this research has focused mainly on prediction model's development, regarding both application and methodology matters such as selecting the most important predictors in the model. The emphasis in the next of chapter is to the validation performance of a prediction model. All prediction models require external validation to check that the model predicts reliably in new data from similar (or even different) settings or populations. In the next chapter, these methods will be externally testing the model that has been developed for prediction of functional upper limb recovery post-stroke using a new data set.

Chapter six

6 External validation

If models are to be adopted for clinical decision making, it is vital that they can accurately predict outcomes in the real world/clinical setting. Prior to implementation in the real world-clinical setting, it is necessary to test the model performance on a new, unseen data set. This process is known as external validation. Within modelling, it is bad practice to advocate models for use when they have only been internally validated, given that model performance is likely to be inflated due to overfitting, given that the model is able to explain the seen or existing dataset. Because of the optimism problem (overfitting) of predictive models, models have worse performance in new patients/subject than expected based on the performance estimated from the development data-set (Harrell et al., 1996; Kwah and Herbert, 2016).

I have previously completed the internal validation processes and identified the four predictors selection models, as described earlier. Therefore, the next step in the model development process is external validation. During the external validation stage, it is also important to assess how good the model prediction is in estimating the errors. This is known as model performance. Here, I compare the external validation performance of the Stepwise logistic regression, LASSO, adaptive LASSO and group of LASSO models on a new dataset.

6.1 External validation study

In this study, the external validation of the developed models was tested by utilising data adopted from an independent randomized control trial of patients who have significant impairment of function in the affected arm post-stroke; more detail in study presented by Lindsay et al. (2014). I extracted a set of predictors from the control group of original data-set. These predictors are similar to the set that was selected by stepwise logistic regression, adaptive LASSO, Group of LASSO and LASSO. Logistic regression of each method was used to re-develop the prediction model of the new data-set.

The following metrics were used to evaluate the externally validity of the model performance. The Brier score was used to assess the overall performance of each of the four models (Harrell et al., 1996). Hosmer and Lemeshow goodness of fit test and calibration plot were used to assess the agreement between the predicted ARAT outcome for each patient and the actual ARAT outcome at three and six months respectively. Discrimination was assessed using: 1) A confusion matrix was used to evaluate the fit of the four models and identify the rate of true positives it classifies as being positive (the sensitivity) and the rate of true negatives it classifies as being negative (the specificity). 2) The discriminative ability of the four models for the upper limb recovery was calculated by measuring the area under the ROC curve and plotted. Finally, in order to investigate the prediction model usefulness in clinical assessment a decision curve analysis (DCA) was evaluated using the net benefit measure(Steyerberg and Vergouwe, 2014; Vach, 2013).

6.2 Baseline characteristics

The Data-set included more than 500 columns of predictors which can be categorised into demographical, historical and clinical measurement variables. The data-set also included the ARAT outcome measures determined at different time points i.e. baseline, 3 and 6 months. A range of (0 to 5 weeks) was at the baseline measurement time point identified for stroke patients between (0-5) weeks of 120 patients (Lindsay et al., 2014). The number of patients is 120 patients at baseline that classified into 73 patients as a control group and 47 of patients as the intervention group. The data set comprises a significant amount of missing -data, and this amount is increasing over the follow-up study. Additionally, more than 80 per cent of values within some variables were constant and thus have been excluded, the remaining predictors in the development and external validation sets are presented in Table 6-1.

*Table 6-1 Summary of the baseline characteristic of predictor variables of both set development** and external validation.*

Predictors	**Odd ratio (OR)**, model development**	**Odd ratio (OR), model validation**
Age	71.3± 11.3	65.46+16.93
Previous stroke	125(83.3%)	
Barthel index Baseline	10.03±5.75	3.51+ 4.61
After Stroke Pain (10) for affected side -Baseline	146(97.3%)	29.78+36.04
NIHSS of arm Baseline	8(5.33%)	3.61+0.63
NIHSS of Leg	101(67.33%)	2.78+ 1.06

NIHSS of Sensory	48(32%)
	22(12%)
		1.15+0.73

Action Research Arm test	20+27.81	13.05+19.81

** the odd ratio of model development comes from the study in chapter four.

6.2.1 Results for external validation of the models at three six months

These results included tests of the ability of the external validity of the four predictions of development models. Four prediction models were assessed based on the result of model's development variable selection process and internally validated, as follows:

1. The equation of Stepwise logistic regression model that used to predict the probability of recovery patients' upper limbs' function after three months is:

$$P= \frac{\boldsymbol{e^{140-0.36(Age)-5.499(NIHSS)+1.7(BI)}}}{\boldsymbol{1+e^{140-0.36(Age)-5.499(NIHSS)+1.7(BI)}}} \tag{6.1}$$

Where:

P: is the predicted probabilities the ARATs' outcome of patients after three months.

NIHSS is the severity of stroke measures.

BI: is the Barthel Index at baseline time.

2. The equation of ALASSO logistic regression model that used to predict the probability of recovery patients' upper limbs' function after three months is

$$P = \frac{e^{3.58-0.56(NIHSS)+0.64(BI)}}{1 + e^{3.58-0.56(NIHSS)+0.64(BI)}} \tag{6.2}$$

Where:

P: is the predicted probabilities the ARATs' outcome of patients after three months.

NIHSS: is the severity of stroke measures at baseline.

BI: is the Barthel Index at baseline time.

3. The equation of GLASSO logistic regression model that is used to predict the probability of recovery patients’ upper limbs’ function after three and six months respectively is

$$P = \frac{e^{0.25-0.036(Age)-0.05(NIHSS)+0.15(BI)}}{1 + e^{0.25-0.036(Age)-0.05(NIHSS)+0.15(BI)}} \qquad (6.3)$$

Where:

P: is the predicted probabilities the ARAT’s outcome of patients after three months.

Age: is age of patient at baseline post-stroke.

NIHSS: is the severity of stroke measures at baseline.

BI: is the Barthel Index at baseline time.

4. The equation of LASSO logistic regression model that used to predict the probability of recovery patients’ upper limbs’ function after three months is:

$$P = \frac{e^{1.01-0.024(Age)-0.048(NIHSS)+0.058(BI)+0.24(previousstroke)+0.007*Pain}}{1 + e^{e^{1.01-0.024(Age)-0.048(NIHSS)+0.058(BI)+0.24(previousstroke)+0.007*Pain}}} \qquad (6.4)$$

Where:

P: is the predicted probabilities the ARATs’ outcome of patients after three and six months respectively.

Age: is age of patient at baseline post-stroke.

Pain: represents patients’ pain at baseline post-stroke.

NIHSS: is the severity of stroke measures at baseline.

BI: is the Barthel Index at baseline time.

The overall performance of models as in the equations (1,2,3,and), the ability of each model to discriminate the differentiation between a patient with the recovery event from a patient without, calibration and ability of the model to improve the decision making procedure (clinical usefulness) were tested (Vach, 2013), as follow:

6.2.2 Performance of recovery functional upper limb prediction model

Due to the Brier score results of each model (Stepwise logistic regression, ALASSO, Group LASSO and LASSO) the Adaptive LASSO has a lower distance between the actual and predicted outcome than other models. ALASSO has better overall performance prediction at three and six months, respectively (Table 6-2).

Table 6-2 Overall performance of recovery functional upper limb of each prediction model.

Prediction Models	**Brier Score of Externally validated prediction models after three months.**	**Brier Score of Externally validated prediction models after six months.**
Stepwise logistic	0.19	0.16
Adaptive LASSO	0.11	0.13
GLasso	0.15	0.31**
LASSO	0.16	0.135

**The Brier score can be accounted for logistic regression model and is the average squared differences between actual outcome (0, 1) and predicted probabilities (ranges from 0 to 1). Where the Brier score was equal to zero that means the model has perfect achievements. If the score was less than 0.25 that indicates good model performance. The model's performance was acceptable according to Brier scores that were (19%, 11%,15% and 16%) at three months and (16%, 13%, 31% and

14%) at six months. A related test to the Brier score is Nagelkerke R2 which is interpretable as the rate of ARATs' outcome variation, which can be clarified by the predictors of the model.

6.2.2.1 Calibration of externally validated prediction models after three and six months

Calibrations of external validation models (Stepwise logistic regression, ALASSO, GLASSO, and LASSO) were deduced using: The Person correlation of goodness fit test and the Calibration plots method.

6.2.2.1.1 Correlation test:

Hosmer–Lemeshow is used to assess the goodness-of-fit χ^2 test. However, because the sample size is small and the Hosmer–Lemeshow test is known to be oversensitive to small sub-group deviations, even if it has good fitting in moderate data-sets it is not suitable in this study. Therefore, to compare the predicted ARAT versus actual ARAT outcomes, the Pearson correlation coefficient was used. (Zhang et al., 2013)

Table 6-3 Results of correlation coefficient of Externally validated prediction models after three months.

Prediction Models	**Correlation coefficients between predicted and actual outcomes after three months.**	**Correlation coefficients between predicted and actual outcomes after six months.**
Stepwise logistic	0.35	0.40
Adaptive LASSO	0.70*	0.62*
GLasso	0.44	0.45
LASSO	0.53	0.50

* highly correlated between the actual and predicted outcomes of ALASSO.

6.2.2.1.2 Calibration plots:

Calibration plots of prediction model are an essential aspect for external validation. These compare the averages of actual outcomes versus predicted outcomes determined from the prediction model. The results of the calibration slope plots were decomposed into results in Table 6-4 and calibration plots of prediction models after three and six months, respectively, in order to display individual model results in a simple way, Table 6-4 and Figure 6-3, Figure 6-2.

Table 6-4 Result of slopes and intercepts calibration of each external validation model.

Prediction Models	**Externally validated prediction models after three months.**				**Externally validated prediction models after six months.**			
	Stepwise	**ALASSO**	**GLASSO**	**LASSO**	**Stepwise**	**ALASSO**	**GLASSO**	**LASSO**
Calibration intercept	-0.72	-0.16	0.29	1.7	-0.43	0.15	0.70	-0.43
Calibration slope	1.79	0.63	1.33	-0.72	1.7	0.56	1.56	1.7

Table 6-4 shows the slope and intercept of the calibration plot of the external validation of four models. Calibration is not close to one and indicates that the model is optimistic. Because the value of the intercept is related to the value of the slope, the intercept automatically changes when the slope changes. In the external validation data, the intercept of each model was (-0.72, -0.16, 0.29 and 1.7) and (-0.43, 0.15, 0.70 and -0.43) for three and six months, respectively. As the results showed, the calibration of stepwise logistic, GLASSO and LASSO models have a poor performance compared to the ALASSO method. The main components of the output of these figures are explained as follows:

The sold line represents the actual model performance that compares the proportion of predicted and actual outcomes of recovery upper limb. A calibration slope of less than one is a sign of the overestimation/overfitting of estimated coefficient, whereas points located above the diagonal line correspond to underestimation prediction. When an intercept of each model is different from 0 that indicates the predicted probabilities are systematically too high (intercept < 0) or too low (intercept > 0). In a sense, the calibration of the intercept represents the term of bias in the prediction model, which is systematic under or over-prediction of probabilities. If both the slope differs from 1 and the intercept differs from 0, the interpretation of the mis-calibration is difficult, because the values of intercept and slope are related. Consequently, the presence of mis-calibration in models has an adverse effect on the model's prediction performance.

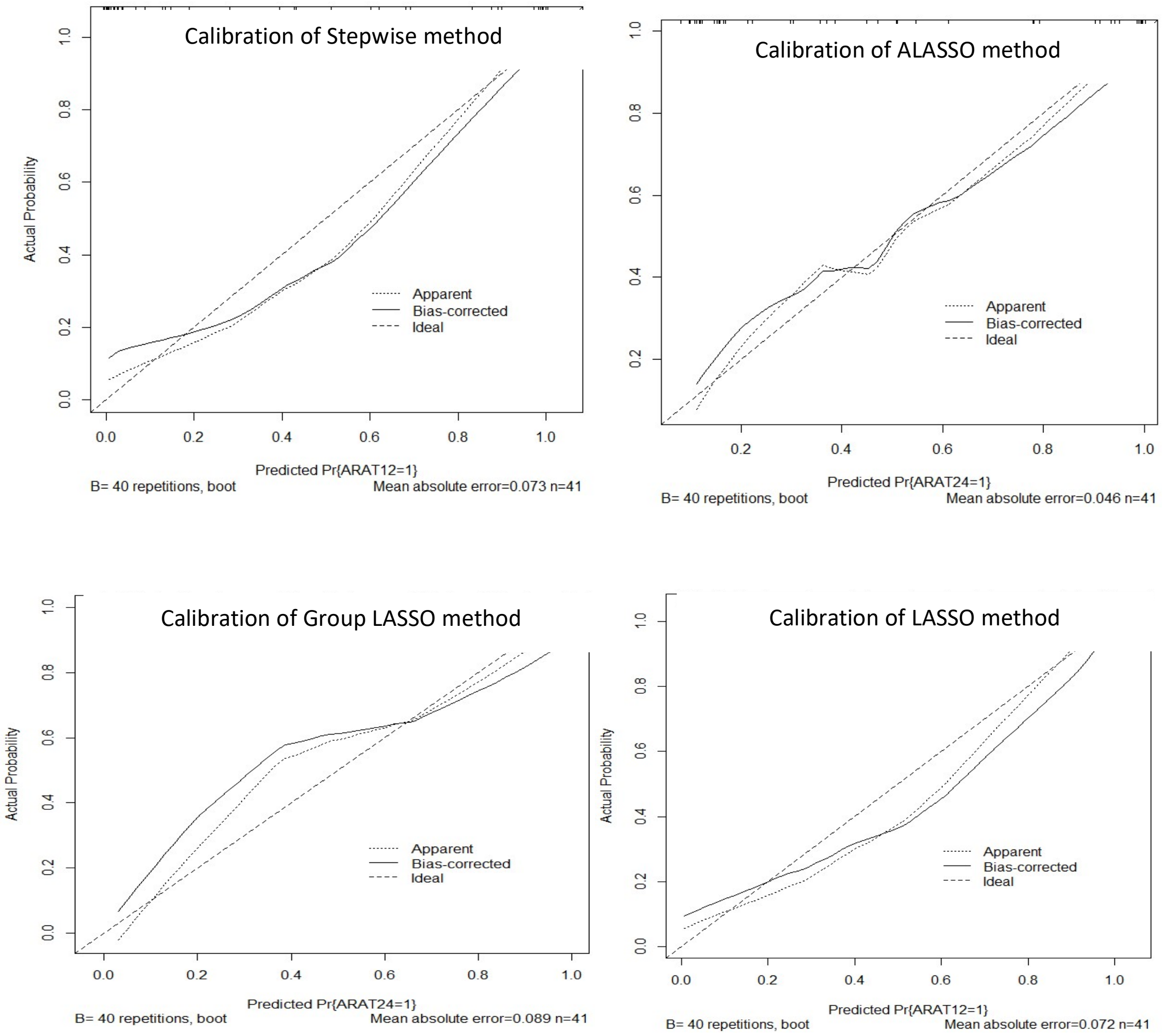


Figure 6-1 Calibration of predicted probabilities recovery of functional upper after three months using four models (stepwise, ALASSO, GLASSO and LASSO) based on the external validation dataset.

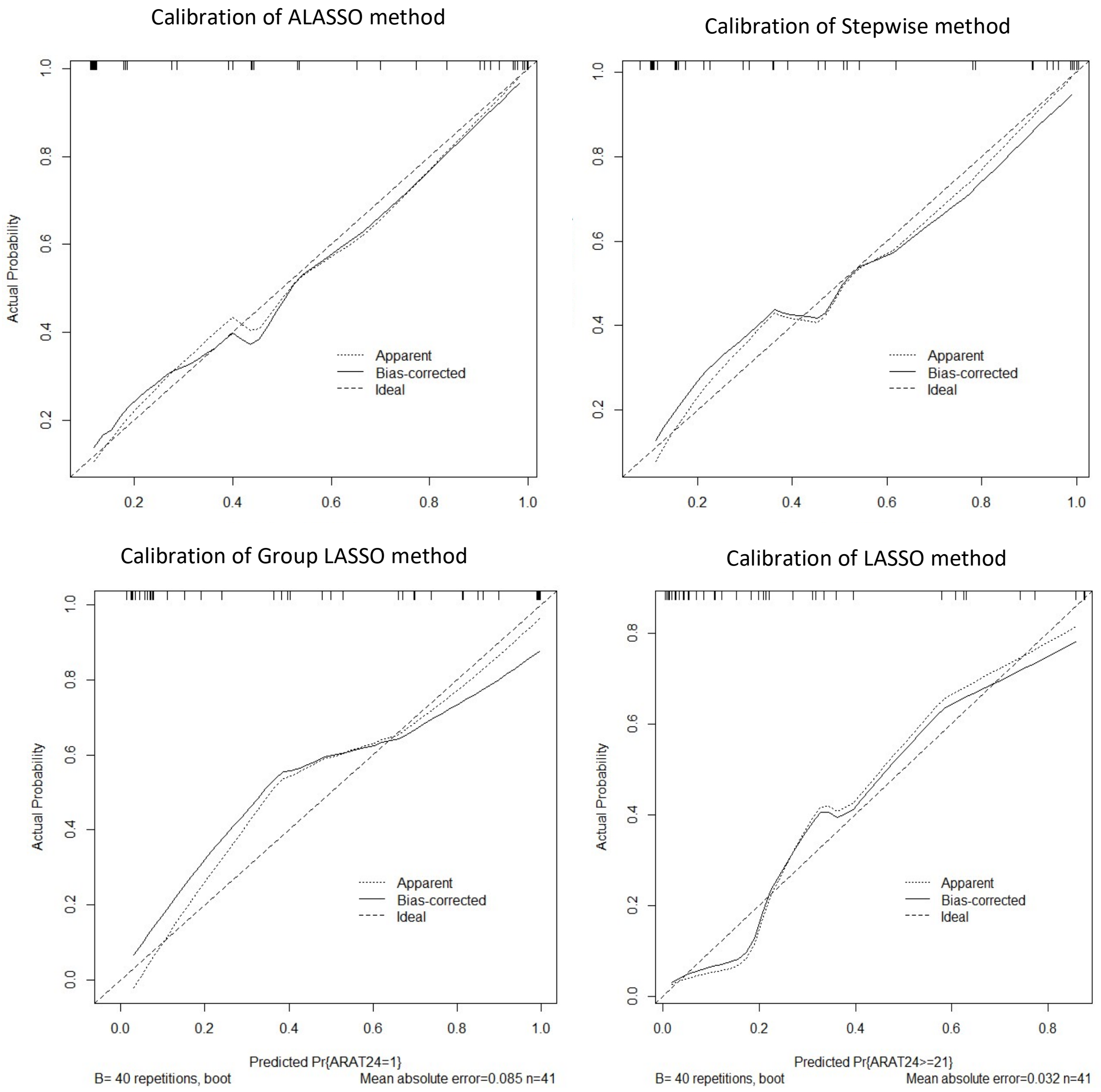


Figure 6-2 Calibration of predicted recovery of functional upper after six months using four models (stepwise, ALASSO, GLASSO and LASSO).

6.3 Discrimination of externally validated prediction recovery models

By using the above four equations of logistic regression models as in section 6.2.1, discrimination was evaluated by finding the predicted probabilities from each model (Stepwise logistic regression, ALASSO, GLASSO and LASSO) for every patient. Sensitivity, specificity and accuracy and the area under the curve of ROC were evaluated for each model at three and six months, respectively,

Table 6-5,Table 6-6. All these steps were deduced using three packages of R (ROCR, AUC and predictABEL).

Table 6-5 sensitivity, specificity, true/ false positive and negative given predicted recovery of functional upper limb of cut-off (ARAT) >=7 after three-month post-stroke of external validation dataset.

Prediction Models	**Stepwise logistic**	**Adaptive LASSO**	**GLASSO**	**LASSO**
No. of predictors	6	2	3	3
No. of True Positive	18	23	23	23
No. of True Negative	8	10	3	7
No. of False Positive	15	8	15	11
No. of False Negative	0	0	0	0
Sensitivity	0.34	0.56	0.16	0.38
Specificity	1	1	1	1
Accuracy	0.63	0.80	0.63	0.73
Positive Prediction Value	1	1	1	1
Negative Prediction Value	0.54	0.74	0.60	0.67
Prevalence	0.44	0.44	0.44	0.44
Area under the curve	0.67	0.78	0.58	0.70

Table 6-6 sensitivity, specificity, true/ false positive and negative given predicted recovery of functional upper limb of cut-off (ARAT) >=7 after six-month post-stroke of external validation dataset.

Prediction Models	**Stepwise logistic**	**Adaptive LASSO**	**GLASSO**	**LASSO**
No. of predictors	**5**	**2**	**3**	**2**
No. of True Positive	**20**	**21**	**21**	**21**
No. of True Negative	**8**	**10**	**3**	**7**
No. of False Negative	**0**	**0**	**0**	**0**
No. of False Positive	**13**	**10**	**17**	**13**
Sensitivity	**0.38**	**0.50**	**0.15**	**0.35**
Specificity	**1**	**1**	**1**	**1**
Accuracy	**0.68**	**0.75**	**0.78**	**0.68**
Positive Prediction Value	**1**	**1**	**1**	**1**
Negative Prediction Value	**0.60**	**0.67**	**0.55**	**0.61**
Prevalence	**0.48**	**0.48**	**0.48**	**0.48**
Area under the curve	**0.69**	**0.75**	**0.57**	**0.68**

A true and false positive rate were evaluated by plotting the area under the receiver operator. The Area under the ROC of (Stepwise logistic regression, ALASSO, Group LASSO and LASSO) were (0.67, 0.78, 058 and 0.69) with confidence interval 95% [(0.49, 078); (0.65, 0.91) ;(0.46, 0.78); and (0.57, 0.86)], respectively. The results of four model's external validations plots show that discriminatory ability of (the Stepwise logistic regression, ALASSO, Group LASSO and LASSO) is not particularly good, reliable with the predictors not explaining much of the difference in the datasets. By contrast, the adaptive LASSO plot shows that the discriminatory ability of this model is good and reliable with the predictors explaining well the variation in the external validation datasets. The ALASSO model's plot shows the effects of the level of severity of stroke and the level of activities to daily living to the recovery of the upper limb of patients; this produced a better performance than the other models (Stepwise logistic regression, ALASSO, Group LASSO and LASSO) Figure 6-3, Figure 6-4.

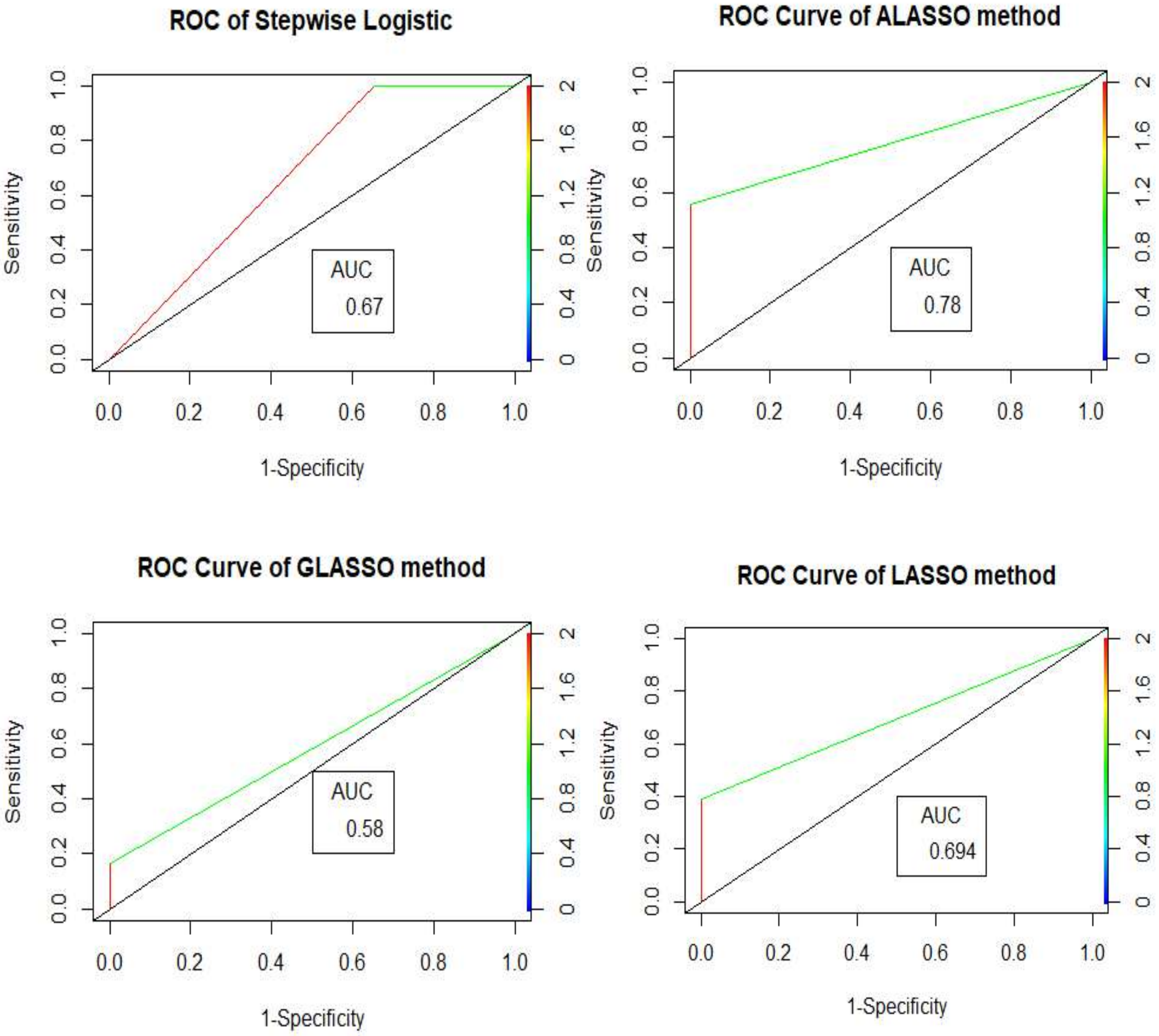


Figure 6-3 Receiver operator curves for the Stepwise logistic regression, ALASSO, GLASSO and LASSO model to predict ARATs' outcome after three months.

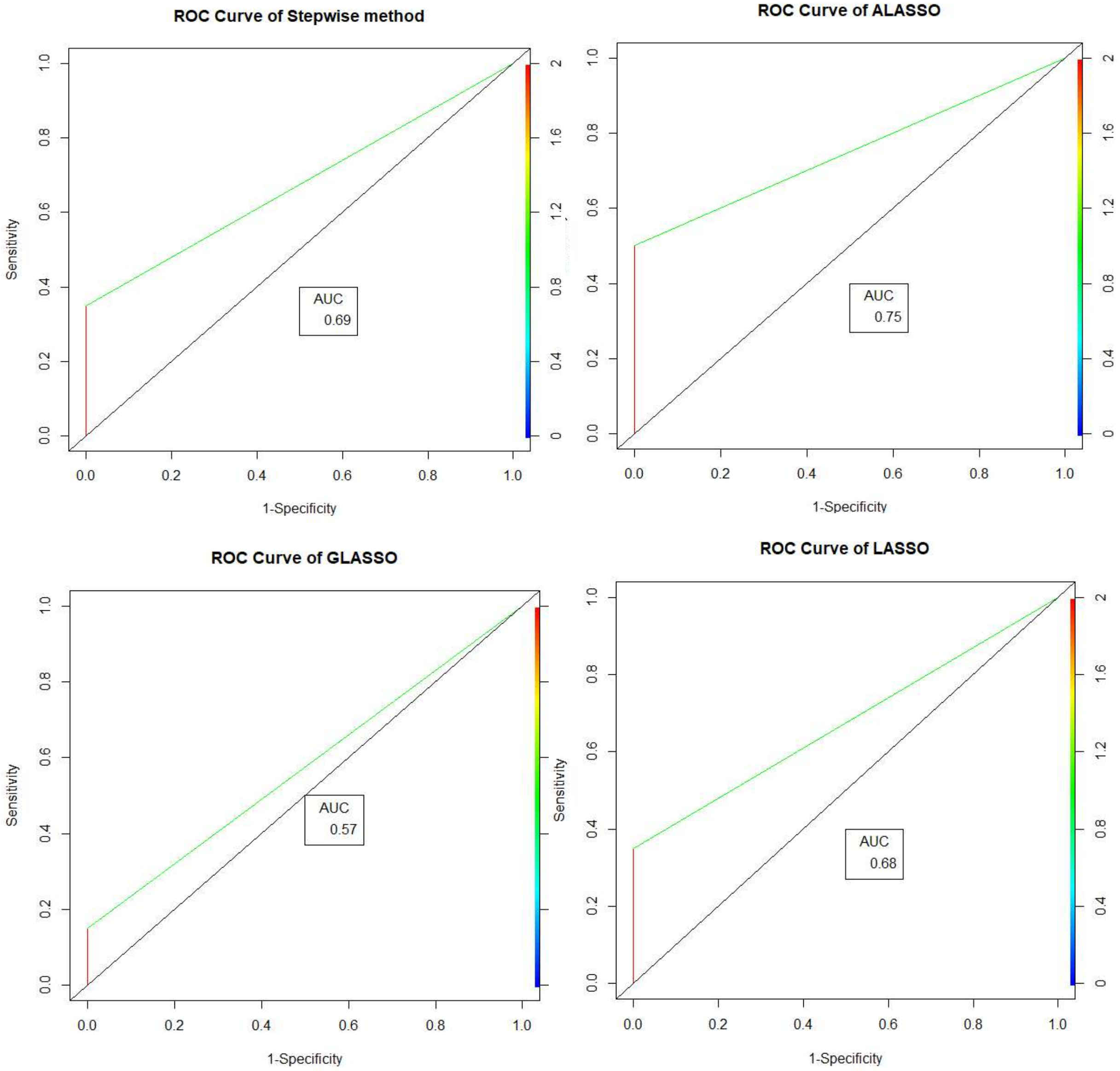


Figure 6-4 Receiver operator curves for the Stepwise logistic regression, ALASSO, GLASSO and LASSO model to predict ARATs' outcome after six months.

6.4 Decision-curve analysis of externally validated prediction models

The aim of this research is to develop a prediction model which can classify patients' likelihood of achieving upper limb recovery and those not likely to achieve recovery to guide rehabilitation programs. Therefore, a cut-off point is required to classify patients as either not being able to recover or being able to achieve recovery so that treatment may be allocated or withdrawn appropriately. At the threshold, the likelihood of improvement, for example reduced impairment because of rehabilitation program therapy, exactly balances the likelihood of no recovery, for example improves the clinical costs-effectiveness. Irrespective of the fact that a prediction model may achieve a good level of calibration and discrimination (sensitivity, specificity and the area under the curve of ROC), these characteristics do not enable the model to assess clinical usefulness (Steyerberg and Vergouwe, 2014; Zhang et al., 2018).

To overcome this limitation, Vickers and Elkin (2006) have proposed a series of decision-analytic measures to summarize the performance of the model in supporting decision making. Additionally, they derived a new tool as a part of decision curve analysis (DCA). This is based on subtracting the rate of all patients identified as false positives from the rate of true positives. The subtraction result is then weighted by using the relation between the false-positive and false-negative results of a prediction model. This tool is called a Net Benefit (NB) that refers to weighting a relative between the two false conditions has a formula as follows:

$$Net\ Benefit = \frac{TurePostiveCount}{n} - \frac{FalsePositiveCount}{n}\left(\frac{p_t}{1-p_t}\right) \qquad (1)$$

Where:

- True- positive count and false positive count represent the number of patients with the true and false positive prediction models results.
- n is the sample size (total number of patients).
- p_t: is where the expected benefit of intervention is equal to the expected benefit of avoiding intervention.

There are two important benefits behind using DCA. First, DCA can be used to compare different types of models. For example, compare results from a predictive model and results from the clinical decision. Secondly, prediction models' benefit in clinical practice can be quantified in a simple way that does not require information on the cost-effectiveness' or how patients perceive their different health states. (Holmberg and Vickers, 2013; Van Calster et al., 2018).

The DCA was used in this section to test the clinical utility of each model and to make comparisons between the Adaptive LASSO performance the other models (stepwise logistic regression, GLASSO and LASSO). The DCA, with NB of each models, was plotted for external valuation after three and six months, respectively, using the functions in R (Zhang et al., 2018). Table 6-7 shows that the Net Benefit results of the four models' external validation tests, which were obtained from a probability threshold of (0.5) for each model.

Table 6-7 The net benefit (NB) results of four external validation prediction models

Methods	**Stepwise logistic**	**ALASSO**	**GLASSO**	**LASSO**	**Treated ALL**
Net benefit of predicting after three months	0.24	0.43	0.10	0.29	<=0.21
Net benefit of predicting after six months	0.02	0.39	0.30	0.0.09	<=0.201

The output of the Table 6-7 appears that the adaptive LASSO prediction, after three months of the external data, is always superior to the other prediction performance (Stepwise logistic regression, GLASSO, and LASSO). At that threshold, the Net Benefit of all treated patients was (0.21) lower that the Net Benefit of ALASSO, which was (0.43

and 0.39) representing predictions after three and six months, respectively. Additionally, the ALASSO has Net Benefit greater than all the other methods. To illustrate the ALASSO superiority over the other models, I need to calculate the difference between Net Benefit of each model and Net Benefit of all-patients-treated. At 0.05 threshold, according to the interpretation given above, this means that one can demonstrate the difference by the following subtraction (0.43–0.21 =0.22). Further, ALASSO has been shown to have 22% higher Net Benefit than when all patients received treatment, which makes our clinical decision based on ALASSO more accurate, hence higher beneficial treatments. In more clinical terms, adaptive LASSO has higher accuracy to exclude patients who might not benefit from rehabilitation (net of false positive), which produce a better more cost-effective clinical decision-making (Holmberg and Vickers, 2013).

The output of the clinical usefulness comparison of ALASSO prediction model's performance with other models were plotted that are shown in Figure 6-6, Figure 6.7, Figure, 6-8, Figure 6-9 and Figure 6-10. The net benefit is plotted against the threshold probability. The "all" line shows the net benefit by treating all patients, and the "none" line is the net benefit for treating none patients.

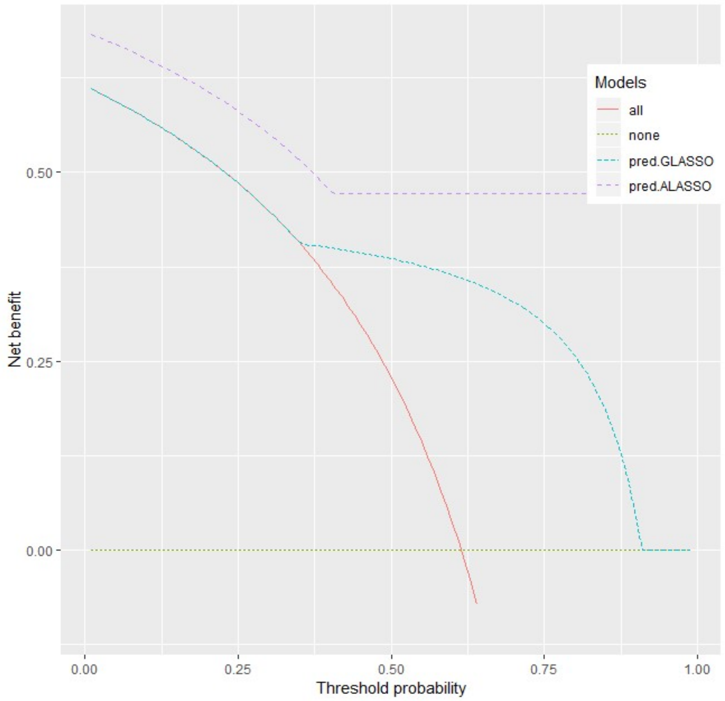


Figure 6-5 Decision curve analysis for the Stepwise logistic regression and ALASSO model for prediction recovery functional upper limb after three months. The two curves are compared to the curves of non-treated and all patients treated.

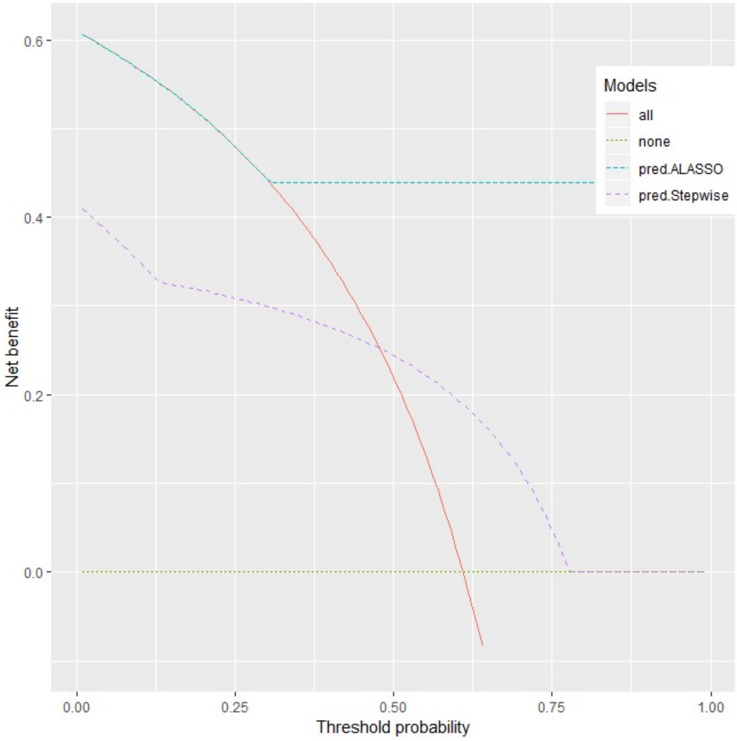


Figure 6-6 Decision curve analysis for the GLASSO and ALASSO model for prediction recovery functional upper limb after three months. The two curves are compared to the curves of non-treated and all patients treated.

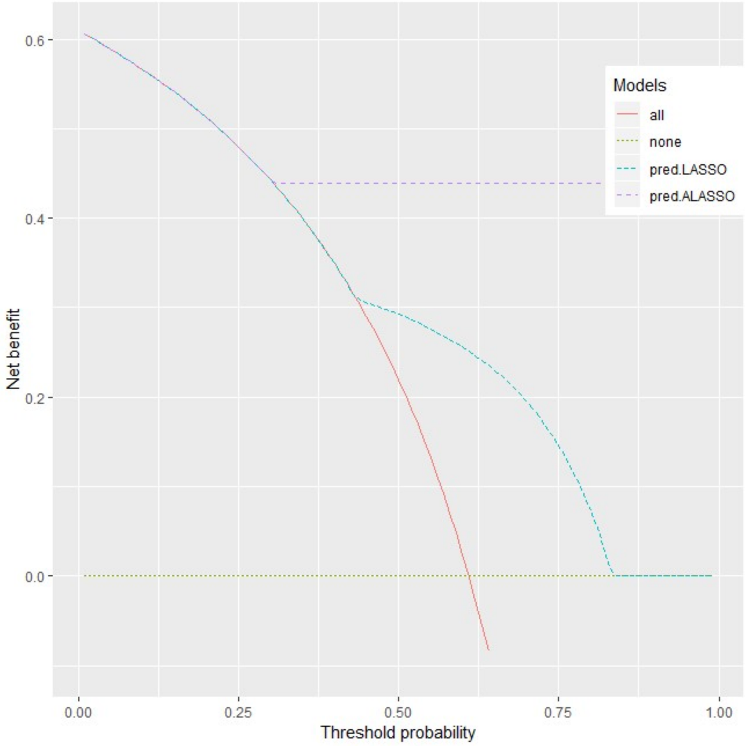


Figure 6-7 Decision curve analysis for the LASSO and ALASSO model for prediction recovery functional upper limb after three months. The two curves are compared to the curves of non-treated and all patients treated.

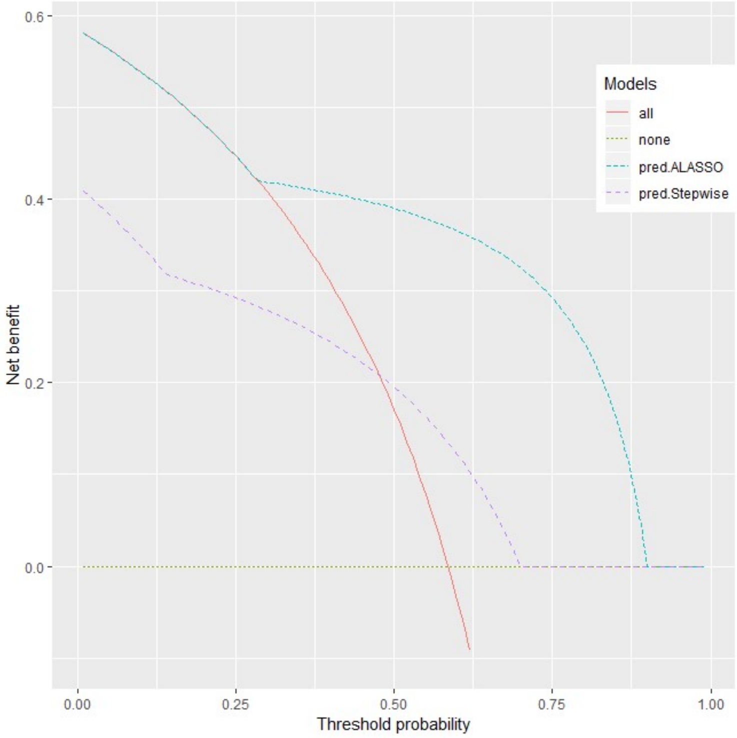


Figure 6-8 Decision curve analysis for the Stepwise logistic regression and ALASSO model for prediction recovery functional upper limb after six months. The two curves are compared to the curves of non-treated and all patients treated.

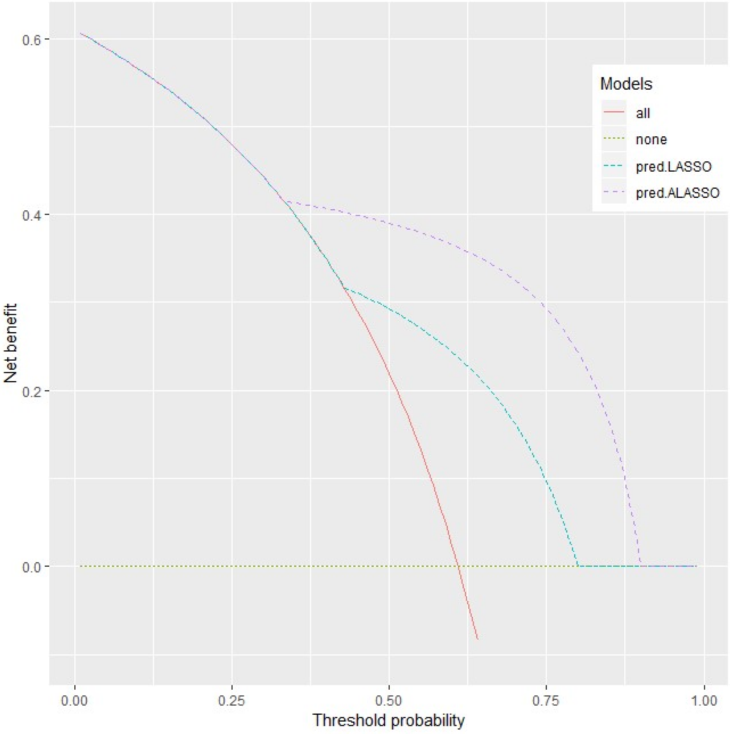


Figure 6-9 Decision curve analysis for the LASSO and ALASSO model for prediction recovery functional upper limb after six months. The two curves are compared to the curves of non-treated and all patients treated.

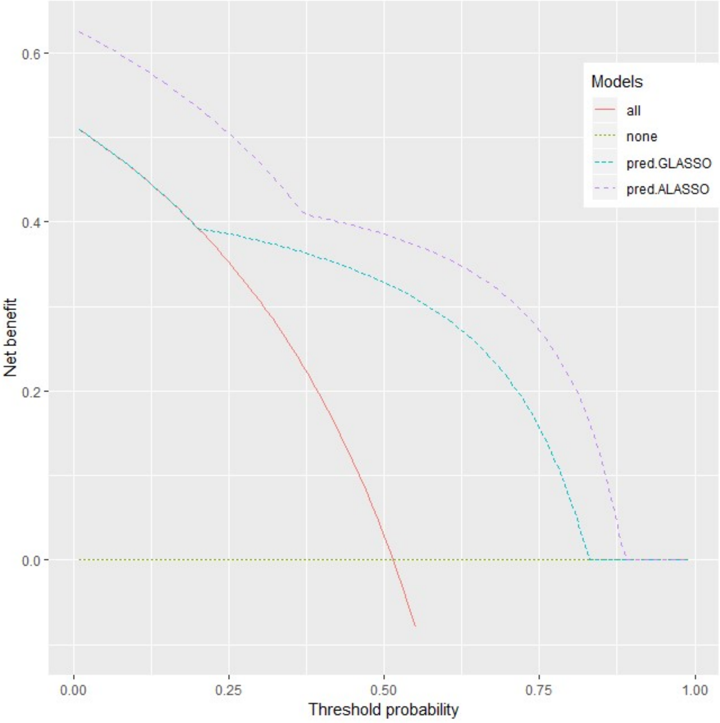


Figure 6-10 Decision curve analysis for the GLASSO and ALASSO model for prediction recovery functional upper limb after six months. The two curves are compared to the curves of non-treated and all patients treated.

6.5 Discussion

This chapter provided a framework to test and compare the external validity of four prediction models developed using classical method (stepwise logistic regression) and three penalised methods (ALASSO, GLASSO and LASSO). A testing process was included to check the overall performance, calibration, discrimination and decision curve analysis of the models.

For the overall performance results of each model, tested using a dataset of a new group of patients, was obtained from the control group of retrospective randomised control trial (40 patients). The model's performance was accepted according to Brier scores Table 6-2. A related test to the Brier score is Nagelkerke R^2 is interpretable as the rate of ARATs' outcome variation, which can be clarified by the predictors of the model.

The R^2 values can be approximated by the difference in the average predicted probabilities of the two groups of patients with different outcomes of ARAT. The ALASSO method was the best model for predicting recovery of upper limb at three and six months, with the difference in average predicted probabilities being (0.66 and 0.56 respectively), Table 6-2 .

Calibration involved comparing the actual ARATs' and predicted ARAT outcomes Table 6-3. Calibration plots for the predicted recovery of functional upper limb of each model are shown in Figure 6-1 and Figure 6-2 . The odds ratios for the overall mis-calibration were from (- 0.72, -0.16, 0.29 and 1.7) and (-0.43,0.15, 0.70 and -0.43) for the four models. ALASSO was best calibrated overall, with intercept of (-0.16) than the other three models. The predictions, after three and six months, explained the actual recovery at best in the ALASSO model (slope of 0.66 and 0.56) and it was at its the worst in the other three models (slopes of 1.79, 0.133 and -0.72) and (1.7, 1.56 and 1.7). The ALASSO model with only NIHSS and Barthel index at baseline calibrated relatively good predicted recovery of upper limb which was better than the other three methods. However, underestimation and overestimation of recovery of upper limb functions at lower and higher predictions were common to all models to some degree.

The discrimination of the external validity of the four models in this study were assessed by area under the curve of ROC. The area under the ROC for each model is classically between (0.6 and 0.85). ROC of ALASSO was 0.88 in the stage of internal validity of the model

ALASSO yielded the best results in sensitivity of about 91%, in the internal validation stage and 56% in the external validation stage. Additionally, the ALASSO's ROC was 0.78 in the two external validity stages (after three and six months), meaning that the model had reasonable capacity to correctly distinguish between patients who would have a higher recovery chance and not. This could relate to stability of ALASSO model's estimated coefficient (Zou, 2006).

The good value of predictions reaching (0.56 of sensitivity) confirms that ALASSO model distinguishes a relatively moderate amount of change in the sample. However, the other three models performed poorly in in the external validity stage. The stepwise logistic, GLASSO and LASSO model had AUC of (0.69, 0.57 and 0.68) and sensitivity of (38%, 15%, and 35%) which are not of practical value. The narrow range of predictions reaching only about (15% to 38% of sensitivity) confirms that these three models are only catching a relatively small amount of change in the sample. Clearly, a big sample size is required to include all the predictors, which are selected in the internal validation stage to check the external validation. These could help to present a useful model practically.

As a result of using the decision curve analysis to evaluate the utility of four prediction models in clinical decisions, the model positively influences our clinical decisions regarding prioritising patients' treatment based on their on-set condition(Vickers and Elkin, 2006). The ALASSO has the net benefit value higher than the other models Table 6-7.

The net benefit, with visualisation in a decision curve, is a simple summary measure to quantify clinical usefulness when decisions are to be supported by a prediction model. If a threshold is clinically well accepted, such as the 50% (representing thresholds for recovery of functional upper limb events), classification tables and its associated measures may be particularly useful.

Finally, the external validity of ALASSO model, developed from the RCTs database, discriminated and classified 'recovery' and 'not recovery' of functional upper limb patients relatively better from all the other methods. This advantage of ALASSO could be of clinical value as its external validity was tested using a different dataset than the one used in its development. On the contrary, using the same dataset used in ALASSO external validity testing, the performance of the other three methods yielded poorer results, giving the superiority to ALASSO.

Chapter seven

7 Developing model

Previously, I have demonstrated how the predictors selection can be improved using the ALASSO instead of stepwise logistic regression, GLASSO and LASSO. I have also shown the performance of each model of selection was achieved in external validation dataset. Additionally, the results showed that the Adaptive LASSO had superiority over the other three methods in predictors selection to predict the recovery of upper limb function in stroke patients, with better external validation than others. Therefore, it is of interest to see if the model can be implemented to explore if it can identify appropriate factors that predict recovery when an intervention is given. Moreover, this study is the first time a modelling method has been used to explore predictors that can emerge if an intervention is used and have demonstrated that this may be possible as "proof of concept".

7.1 Dataset

Data was adopted as the only interlineation group of retrospective study of stroke patients surviving with a significant impairment of the arm function(Lindsay et al., 2014). The intervention group includes 70 patients and different type of predictors demographic and clinical measure at baseline, such as Barthel Index and Modified Rankin Scale. Additionally, data-set includes the three times of outcomes measure for Action Research Arm Test (ARAT) at baseline, three months and six months post-stroke. These associated variables are shown in Table 7-1.

Table 7-1 Predictive characteristics of a studied prediction model of treatment.

series	Demographic Predictors	Mean ± Standard deviation
1.	Spasticity identified	14.85 ±8.9
2.	Stroke to inject	18.88 ±9.65
3.	Age	68.17±14.87
4.	Hemiplegic side Yes No	 (33) 68.8% 15(31.3%)
5.	Infarct or Haem 0 1 2	 (9) 18.8% (29) 60.8% (10) 20.8%
6.	Thrombolysed Yes No	 (39) (81.3%) (09) (18.8%)
7.	Area of Damage 0 1 2 3	 (5) 10.4% (11) 22.9% (16) 33.3% (16) 33.3%
8.	Previous stroke No yes	 (29) 60.4% (19) 39.6%
9.	Total of National Institute of stroke scale	16.27±6.07
10.	National Institute of stroke scale Arm • Drift. • Some effort against gravity. • No effort against gravity. • No movement.	 (1) 2.1% (1) 2.1% (15) 31.3 (31) 64.6
11.	National Institute of stroke scale leg • Drift. • Some effort against gravity. • No effort against gravity. • No movement.	 (5) 10.4% (11) 22.9% (16) 33.3% (16) 33.3%
12.	National Institute of stroke scale sensory • Normal • Mild-to-moderate sensory loss • Severe to total sensory loss	 (8) 16.7% (22) 45.8% (18) 37.5%
13.	National institute of stroke scale inattention • No abnormality • Visual, tactile, auditory, spatial, or personal inattention	 (16) 33.3% (16) 33.3%

	- Profound hemi-attention or extinction to more than one modality	(16) 33.3%
14.	Active move baseline Yes No	 (37) 77.1% (11) 22.9%
15.	Barthel baseline	2.94±4.64
16.	Pain baseline	9.02±20.73
17.	Functional arm scale	3.52±1.90
18.	Modify ranking scale	3.79±1.32
19.	Length of stay	60.44±28.59
20.	DC Destination	2.19±1.67

21.	Family care
	Yes
	No
		
		20(41.7);
		28(58.3)

22.	Tardieu at base line Yes No	 (36)75% (12)25%
23.	Range of movement lost Yes No	 (24) 50% (24) 50%
24.	Baseline Elbow Flexion Maximum Strength Best base	1.12±2.52
25.	Baseline Elbow Extension Maximum Strength Best base	0.66±1.88
26.	Baseline GRIP Maximum STRENGTH BEST	0.75 ±2.33
27.	Wrist Flexion Maximum Strength BEST	1.48±2.32
28.	Wrist Extension Maximum Strength Best	1.09±1.91
29.	Elbow Flexion Maximum Strength Best	3.56±5.10
30.	Elbow extension Maximum Strength Best	2.64±3.72
31.	GRIP Maximum Strength Best	2.95± 5.8M
32.	Movement Elbow Mean Velocity Slow at baseline	30.93± 20.07

7.2 Developed model

An overview of the developed model method has been presented previously to include the pre-processing (handling missing data), describing the predictor's characteristics and checking multicollinearity. Developing a process of prediction included two models, one to predict the early use of botulinum toxin in post-stroke spasticity after three months and a second model after six months. The terms model

3m and model 6m were used to simplify the way of presenting in the next sections. The process was followed that includes a few steps, as follows:

7.2.1 Predictors selection

Predictors selection using ALASSO was preceded by finding the optimum value of penalty (λ). The penalty was identified using two steps: inverse ridge regression coefficients were used for each variable as their weight in adaptive LASSO. Then, estimating the value of tuning parameters, which corresponds to the minimum value of binomial deviance and then the final selection of predictors. Tuning parameter was estimated using 10-fold cross-validation method. This method determines the optimal values of penalty that represent the solution of predictors selection. The optimal selection is located between the first and the second dash line as in Figure 7-1 that is corresponding to the minimum value of deviance. The optimal selection shrinks unrelated predictors from (32) to four and eight predictors in model 3m and model 6m respectively.

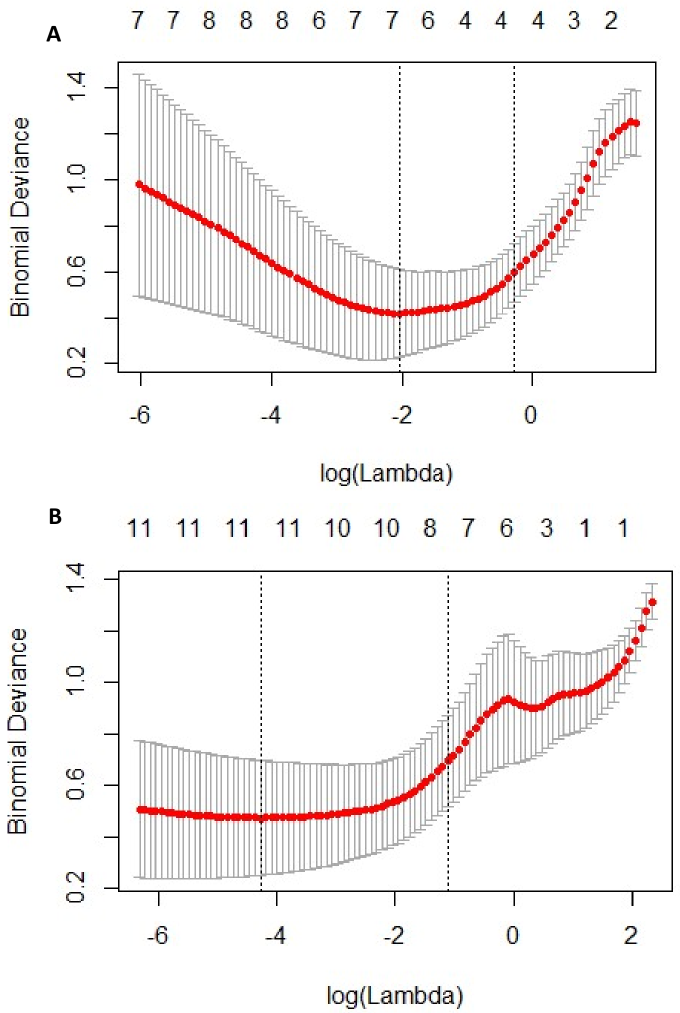


Figure 7-1 Plot of the deviance cross-validation as a function of the penalty parameter λ to determine the estimated optimal value of tuning parameters estimated of ALASSO. The dotted line on the left side corresponds to the λmin specification. The second line is a λ1se specification. A represents the results of model 3m and B represents the results of model 6m.

7.2.2 Results

7.2.2.1 Test multi-collinearity

The VIF was taken to check the multi-collinearity level between predictors.

Table 7-2 Results of multicollinearity test.

Series	Predictors variable	Variances Inflation Factors
1.	Age	3.7
2.	NIHSS of Arm	2.88
3.	Pain Baseline	2.26
4.	Tradieu Baseline	2.39
5.	Active Movement	5.7
6.	Barthel Index	6.54
7.	Wrist Flexion Maximum STRENGTH BEST	20.03*
8.	Elbow Extension Maximum STRENGTH BEST	45.79*
9.	Movement Elbow baseline mean velocity slow	2.4
10.	GRIP Maximum STRENGTH BEST	20.22*
11.	Functional Arm Scale	7.36
12.	Modified Ranken scale	7.38
13.	Stroke to Inject	2.06
14.	Hemiplegic Side	4.06
15.	NIHSS of Inattention	5.12
16.	Infarct or Haem	3.48
17.	NIHSS of Sensory	3.13
18.	Total NIHSS	10.08*

19.	Range of Movement Lost	2.3
20.	Baseline Elbow Flexion Maximum Strength BEST	10.88*
21.	Thrombolysed	4.27
22.	Area of Damage	3.7
23.	Previous Stroke	1.88
24.	Elbow Flexion Maximum Strength BEST	30.96*

Table 7-2 shows that there is a high level of multi-collinearity test among predictors that have values of (VIF>=10).

7.2.2.2 Predictors selection

The two models developed for predicting early use of botulinum toxin in post-stroke spasticity based on ARAT outcome are given in section 7.2). Thirty-two predictors are included in both models and the results of selection of the ALASSO method were four and eight of related predictors for model 3m and 6m respectively. ‘Active move base line’ and ‘Elbow Maximum Strength Best’ were significant predictors in both models. Interestingly, the variables ‘Thrombolysed, National Institute of Stroke Scale Arm, Family Care and Tardieu at base line were not included in model 3m but were retained in model 6m. Range of movement lost was retained in the both models, but it has approximately twice the negative effect in the model 3m than the model 6m. This means, an increase in the ‘range of movement lost’ of a patient will have 50% less benefit from treatment after three months than the treatment after six months. Additionally, the result ALASSO selection predictors found the set of predictors of model 3m appeared within the model 6m set predictors. This would suggest that, this set of predictors are significantly important for the development of a prediction model of intervention of spasticity.

Table 7-3 Odd ratio of estimated regression coefficient using ALASSO for predicting the benefit of intervention based on the outcome of ARAT after three and six months.

Series	Predictors	Odd ratio of the coefficient (3 months)	Odd ratio of the coefficient (6 months)
1.	Intercepts	-1.3095	0.1533
2.	Active move base line	1.4771	1.127
3.	Modify ranking scale	-0.3165	-1.1277
4.	Range of movement lost	-0.7223	-0.4357
5.	Wrist Extension Maximum Strength best	1.008	1.5816
6.	Thrombosed	0	1.0420
7.	National Institute of Stroke Scale Arm	0	-0.0728
8.	Family care	0	-0.9752
9.	Tardieu at base line	0	-0.629

7.2.2.3 Calibration model

In section 7.2 two prediction models (model 3m and model 6m) were developed based on the ARAT outcome. The calibration of each model was internally validated by plotting the mean observed ARAT outcome and the mean predicted from the model.

To do this, the predicted ARAT score was calculated for each patient in the treatment group. Calibration plots for the configuration of two ALASSO models evaluated are shown in Figure 7-2. The ALASSO over predicted the recovery of early used botulinum toxin in post-stroke spasticity patients in the lower ranges of intervention but under predicted the recovery in the higher ranges of intervention.

There are differences between plots from the actual and predicted ARAT outcome Figure 7-2. The predicted ARAT outcome did not lie close to the actual outcome of ARAT. Overall, both models have poorly calibrated.

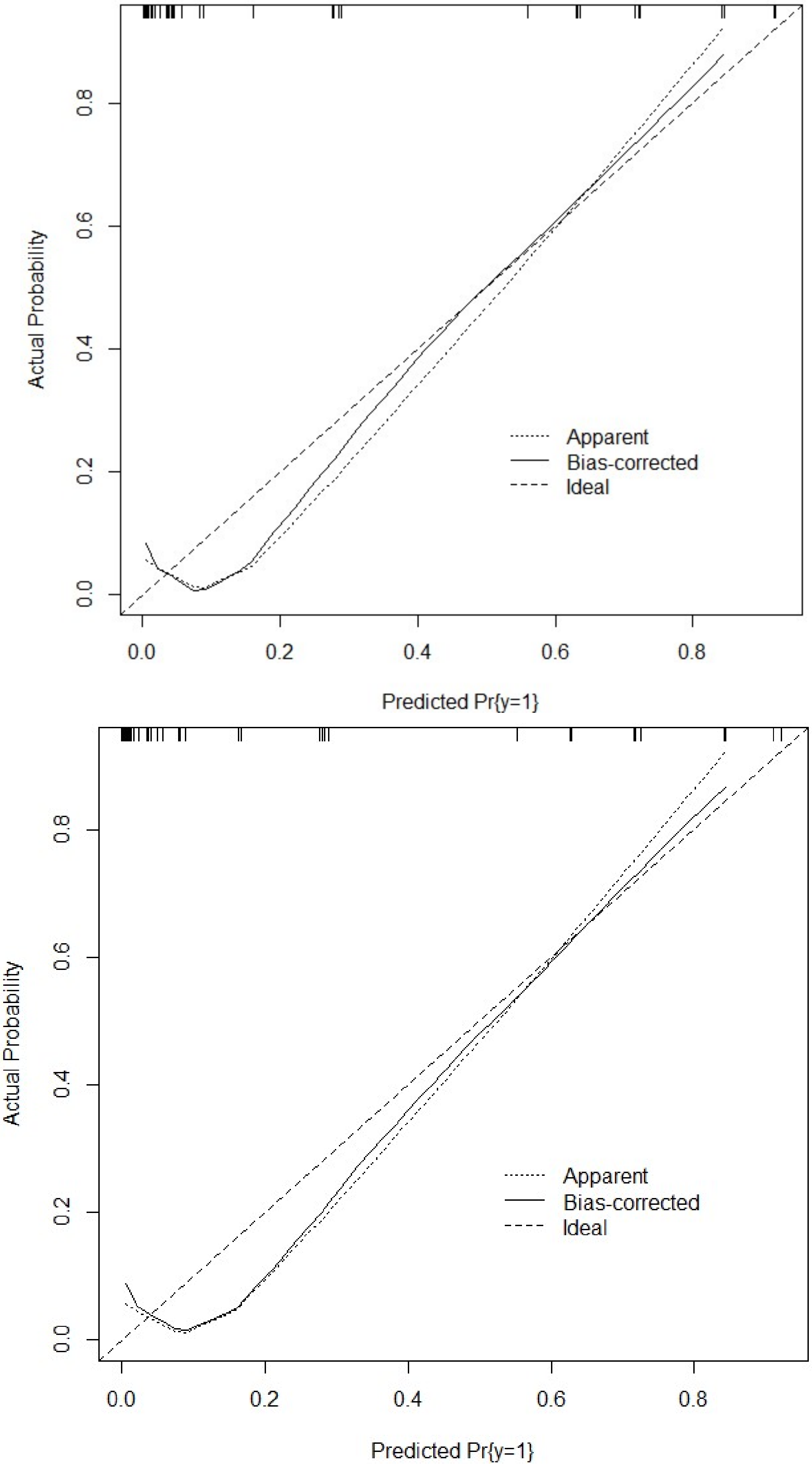


Figure 7-2 Calibration plot of prediction early use of botulinum toxin in post-stroke spasticity after three and six months

7.2.2.4 Discrimination model

There appears to be good discrimination between the patients after intervention when the models are fitted using ALASSO. Similar discrimination was noticed of both models Table 7-4 . The internal validation showed that the models discriminated reasonably well with average C-statistics across imputed datasets of 0.833 and 0.828 for the models 3m and model 6m respectively. These values suggest slightly better discrimination of the model 3m using four predictors in the other treatment group, whereas the model 6m performed slightly less in the other treatment group using eight predictors. Sensitivity and specificity were calculated using the area under the ROC curve. The area under ROC curves for the ALASSO tested was estimated using (ROC package). These are shown in Table 7-5 and Figure 7-3 below.

Table 7-6 Sensitivity, specificity, true/ false positive and negative given predicted recovery of the functional upper limb of cut-off (ARAT) >=7 after six-month post-stroke of external validation dataset.

Prediction Models	**Model 3m is to predict the early use of botulinum toxin in post-stroke spasticity after three months.**	**Model 6m is the early use of botulinum toxin in post-stroke spasticity after six months.**
No. of predictors	4	8
No. of True Positive	34	29
No. of True Negative	9	12
No. of False Negative	1	4
No. of False Positive	4	3
Sensitivity	0.97	0.90
Specificity	0.69	0.75
Accuracy	0.89 (0.77,0.96)	0.85 (0.72,0.94)
Positive Prediction Value	0.89	0.87
Negative Prediction Value	0.90	0.80
Prevalence	0.72	0.66
Area under the Curve	0.83	0.828

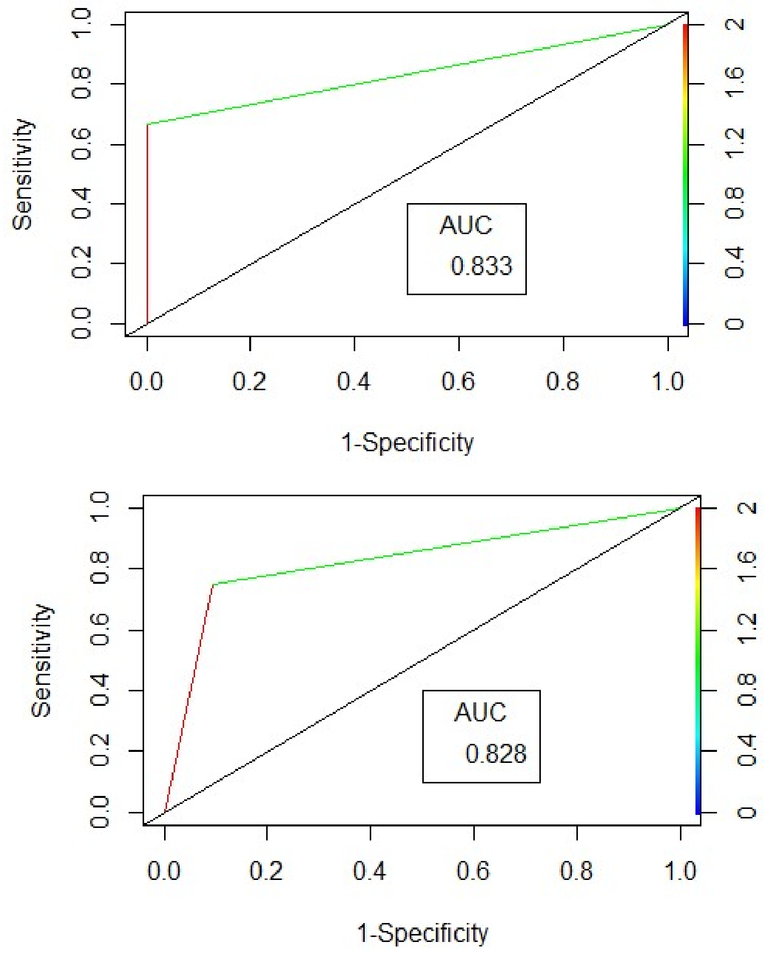


Figure 7-3 ROC Curve of international validation of prediction early use of botulinum toxin in post-stroke spasticity after three months.

7.3 Discussion

Two prediction models developed for patients with spasticity could be used to aid treatment decisions, by potentially identifying patients that could receive a botulinum toxin A (BoNT-A), identifying patients suitable for future clinical trials or off-study treatments.

This chapter used ALASSO modelling to develop a prediction model for spasticity intervention that allows individualised predictions and identified methodological issues when using clinical trials data for this purpose. The important findings and limitations are now discussed.

7.3.1 Summary and comparison to previously published model

A main aim of this chapter was to identify the effective predictors on the early use of botulinum toxin A in post-stroke spasticity and to develop prediction models of the treatment group of RTCs of (Lindsay et al., 2014). In doing so, the models have been built using a different process versus the previous studies that used the backward, forward and stepwise methods of variable selection (Leathley et al., 2004; Moura et al., 2009; Opheim et al., 2015). These methods suffer from lack of stability and are influenced by small sample size relative to large numbers of predictors (Tibshirani, 1996).

Another reason for the difference in model's selection is that previous studies suffer from methodological shortcomings in developing models process. For example, there is no previous study which undertook the internal validation test during the development model such as either using sub-sampling methods (cross-validation) or resampling method (bootstraps) (Steyerberg and Vergouwe, 2014). The authors tested predictors in univariate analysis and selected the significant predictors and applied multivariate logistic methods. Additionally, authors did not deduce the multicollinearity test among the predictor's combinations. As mentioned previously, the existence of multi-collinearity would be a challenge for using traditional methods, particularly in small sample data sets; which could risk the selection of inappropriate or confounded predictor variables (for example a latent variable).

In contrast to this, the two models developed used ALASSO, which is not affected by the multicollinearity level and small sample size relative to large numbers of predictors. Furthermore, the internal validation was undertaken using the calibration and discrimination methods. Previous studies were concerned with identifying the important predictors for predicting the presence of spasticity in the upper and lower limb post-stroke (Moura et al., 2009; Sunnerhagen, 2016). However, in this study, prediction models were developed to predict the recovery from spasticity using of botulinum toxin type A in post-stroke spasticity.

ALASSO in model 3m confirmed the importance of range of movement lost and modify ranking scale as predictors that have negative effect on recovery of spasticity. The active move baseline also was identified and max strength best that have the positive effect on the recovery of spasticity. NIHSS was not a significantly associated predictor to the intervention of spasticity in the model; there was evidence to support these results of NIHSS from the previous study (Opheim et al., 2015). However, it appeared as a significant predictor in the model 6m.

The calibration and discrimination represented the internal validation of both developed models. For calibration, both models were poorly calibrated. Discrimination of both models showed good performance as shown by the AUC of 0.833 and 0.823, the sensitivity of 97% and 90%; and specificity 69% and 75%.

7.3.2 Limitations

Pre-processing data was used to verify the consistency of the predictors and the outcome before trying to build prediction models. However, the dataset was not considered and designed specifically for this study's aims. Important predictors were not defined prospectively (for example Sensorimotor function using the Fugle Meyer test(Opheim et al., 2015). Models performance are usually somewhat optimistic when estimated internally based on sub-sampling and resampling processes, and therefore model discrimination is possible to be even lower when assessed in external data. This means that internal validation is not enough to evaluate.

7.4 Conclusion

This chapter demonstrated how ALASSO could be used to identify the useful predictors and develop a prediction model for recovery by intervention spasticity. New predictions models were developed and internally validated, but due to its excellent discrimination, a new dataset is likely required for external validation before they can be used for practice. Finally, beyond an easy prediction method, the result of this study advises that focusing treatment on the more important predictors is possible to improve recovery.

Chapter eight

8 Discussion and conclusion

8.1 General discussion

Predicting outcome(s) of upper limb function recovery post-stroke is complex but essential and vital as it informs patients and their families, as well as stroke case-manager/clinical decision-maker, about the patients' prognoses and rehabilitation program plan. This project has used modern and traditional methods to examine prediction factors (predictors) and develop different models for patients with upper limb impairment post- stroke. The focus was to study the cut-off point of the response variable (ARAT). The next step was to apply classical and penalised methods to identify candidate predictors, which are collected routinely, and to use these to build and validate a recovery-prediction model. These steps were performed using datasets from secondary anonymised datasets of two previous RCT studies; one dataset was used to build the model, and the other used to validate it. It is worth mentioning that both studied were double blinded studies which indicate a more robust approach.

In chapter one, the literature review was undertaken to identify and investigate published studies on recovery-prediction models. Furthermore, this review included studies of all potential predictors of upper limb recovery post-stroke. Searching studies databases were conducted using the EBSCO interface, which was limited to studies conducted within the past thirty years of databases search date. From the literature, six prediction models were found (Feys et al., 2000a; Hendricks et al., 1997; Kumar et al., 2016; Kwah et al., 2013; Nijland et al., 2010b; Stinear, 2010). It has been shown that very few prediction models exist for upper limb recovery in severe cases post-stroke. Additionally, these models are still mis-categorising some patients. This could be due to the fact that five of the six models did not undergo external validation.

The sixth model was for arm recovery (the proportional recovery model) which has been externally validated. However, it does not give a good prediction in all patients with stroke, as has been suggested from development and validation studies. This model is limited and appears to predict outcomes well with less severe stroke patients. In this current study, external validation has been done in all stroke patients in the control group of RCT. The validation study was conducted on a completely independent sample to the development sample.

The cut-off related to the ARAT's score (dependent variable) in the prediction model was downgraded to seven instead of ten, which was used in the previous studies (chapter three). In this study, 18 patients with ARAT score (<10) were displayed based on their total score of each sub-group within the ARAT score to modifying using bar charts. It has been noticed that 11 of 18 patients recovered after three months, even though these patients, according to the existing models, could be overlooked because they would be categorised as non-recovery patients. Additionally, the patients with total ARAT score of nine or less can incompletely perform some of the easy tasks in the grasp/grip subgroups tests. To ensure that such patients are not overlooked, our decision to reduce the cut-off point to seven was made. Moreover, it might help to make a balance between the cost-effectiveness and interventions of patients, as a lower cut-off point would mean a higher workload on clinicians.

Predictors selection methods of the prediction model for patients with upper limb impairments were studied. The study involved penalised and traditional methods. Uniquely, this study was the first in using penalised methods in model development process of prediction recovery of upper limb function post stroke. The idea behind using the penalised methods is to avoid the impact of the issues of multicollinearity, number of predictors and sample size. It was found that the adaptive LASSO has the superiority in the selection predictors, developing and internal validation model compared with the other penalised method (LASSO and Group of LASSO) and traditional methods (stepwise logistic regression).

Validation is an essential part of the modelling process and therefore external validation methods have been applied to the selected models in this thesis (chapter five). The search of the literature revealed that very few prediction models of the recovery of upper limb post-stroke patients is externally validated, possibly due to the lack of guidance, i.e. from expert statisticians, on suitable validation methods, or possibly due to the lack of appropriate datasets to test external validity.

An external validation study was therefore undertaken to assess methods of externally validating a prediction model and to assess four methods for selection predictors from the external validation dataset. The decision analysis curve method was also applied to present the usefulness of each model in practice. It was found that the ALASSO has the superiority in the external validation via calibration plots and discrimination at three and six months. Additionally, ALASSO has the net benefit compared with the other penalised method (LASSO and Group of LASSO) and traditional methods (stepwise logistic regression). Interestingly, ALASSO success could be attributed to having the advantages of oracle properties of regressions' coefficients, which have proved to be a consistent method of predictors selection(Zou, 2006).

The ALASSO modelling method was used to develop a prediction model that can be used to identify appropriate factors that predict recovery when an intervention is given (chapter six). This study is the first time a modelling method has been used to explore predictors that can emerge if an intervention is used, hence - proof of concept. ALASSO modelling method was thought to be appropriate for this scenario; it was therefore employed to attempt to develop a prediction model which included only treatment groups from RCTs. It is essential to mention, this study requires steps to be taken to test its external validation, which was not performed in this project due to lack of appropriate dataset.

8.2 Research limitation

The main limitation in this thesis was the properties of the data used to develop this study prediction model. For example, some of predictor variables chosen by the model during the development/internal validation stage were not available within the dataset used for the model during the external validation testing stage. Due to the time constraints, the author has not been able to get the same predictors in the data set. Specifically, one of step in this study was to test the external validity using a proper and large enough data-set, with specific properties. Therefore, we requested the data from the Virtual International Stroke Trials Archive (VISTA®); however, we receive a dataset that was not suitable to be used in this study. Obtaining more homogenous data may give a better chance to check the external validation of models. Furthermore, the researchers could not develop a model using sub-group of cluster analysis results. This study used secondary collected data that is not specifically collected for the purpose of model development.

8.3 Conclusion

Prediction stroke research, particularly in the prediction of recovery of upper limb function post-stroke modelling, is still a challenging area that requires more methodology research to improve the models being developed and validated. The aim is to provide useful models that will be implemented in clinical practice, and ultimately, improve patient outcomes and the efficiency of health care delivery.

Though many issues remain, this thesis has contributed toward improvements in the prediction modelling field through application and methodological development. The use of the penalised method (adaptive LASSO) will hopefully improve the development, evaluation, presentation and approval of robust prediction models in the coming years, adjusted for predictors.

8.4 Future works

Further studies could be deployed that researchers can use in the framework of developing a prediction model of recovery of upper limb function post-stroke. First, modelling of sub-group of the cluster analysis results could help in reducing the heterogeneity of the predictive value of the model. Therefore, researchers could focus more on merging the LASSO family methods with cluster analysis models. Second, regarding the tuning parameters, researchers could use another method instead of cross-validation to estimate the tuning parameters, such as BIC. Third, models that were developed in this research in chapter six require more investigation in term of external validation based on the prospective data set.

On the other hand, the methodology of this research could be followed to develop a prediction model in the lower limb research, i.e. to make predictions about gait function of a lower limb of patient’s post-stroke.

9 References

Ackerley, S. J. & Stinear, C. M. 2010. Stimulating stimulation: can we improve motor recovery following stroke using repetitive transcranial magnetic stimulation? *Physical Therapy Reviews,* 15**,** 302-308.

Aggarwal, C. C. & Reddy, C. K. 2013. *Data clustering: algorithms and applications*, CRC Press.

Alexopoulos, E. C. 2010. Introduction to Multivariate Regression Analysis. *Hipokratia,* 14**,** 23-28.

Altman, D. G., Vergouwe, Y., Royston, P. & Moons, K. G. 2009. Prognosis and prognostic research: validating a prognostic model. *Bmj,* 338**,** b605.

Amarenco, P., Bogousslavsky, J., Caplan, L., Donnan, G. & Hennerici, M. 2009. Classification of stroke subtypes. *Cerebrovascular Diseases,* 27**,** 493-501.

Androulakis, E., Koukouvinos, C. & Vonta, F. 2014. Tuning parameter selection in penalized generalized linear models for discrete data. *Statistica Neerlandica,* 68**,** 276-292.

Arboix, A., García-Eroles, L., Comes, E., Oliveres, M., Balcells, M., Pacheco, G. & Targa, C. 2003. Predicting spontaneous early neurological recovery after acute ischemic stroke. *European Journal of Neurology,* 10**,** 429-435.

Arthur, D. & Vassilvitskii, S. k-means++: The advantages of careful seeding. Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms, 2007. Society for Industrial and Applied Mathematics, 1027-1035.

Ashford, S., Slade, M., Malaprade, F. & Turner-Stokes, L. 2008. Evaluation of functional outcome measures for the hemiparetic upper limb: A systematic review. *Journal of Rehabilitation Medicine,* 40**,** 787-795.

Aziz, N. A., Leonardi-Bee, J., Phillips, M., Gladman, J. R. F., Legg, L. & Walker, M. F. 2008. Therapy-based rehabilitation services for patients living at home more than one year after stroke. *Cochrane database of systematic reviews (Online)***,** CD005952-CD005952.

Baird, A. E., Dambrosia, J., Janket, S. J., Eichbaum, Q., Chaves, C., Silver, B., Barber, P. A., Parsons, M., Darby, D., Davis, S., Caplan, L. R., Edelman, R. E. & Warach, S. 2001. A three-item scale for the early prediction of stroke recovery. *Lancet,* 357**,** 2095-2099.

Balaban, B., Tok, F., Yavuz, F., Yaşar, E. & Alaca, R. 2011. Early rehabilitation outcome in patients with middle cerebral artery stroke. *Neuroscience Letters,* 498**,** 204-207.

Barnes, M. P., Dobkin, B. H. & Bogousslavsky, J. 2005. *Recovery after stroke*, Cambridge University Press.

Barra, J., Oujamaa, L., Chauvineau, V., Rougier, P. & Pérennou, D. 2009. Asymmetric standing posture after stroke is related to a biased egocentric coordinate system. *Neurology,* 72**,** 1582-1587.

Beebe, J. A. & Lang, C. E. 2009. Active range of motion predicts upper extremity function 3 months after stroke. *Stroke; a journal of cerebral circulation,* 40**,** 1772-1779.

Breiman, L. 2001. Random forests. *Machine learning,* 45**,** 5-32.

Brott, T., Marler, J. R., Olinger, C. P., Adams, H. P., Tomsick, T., Barsan, W. G., Biller, J., Eberle, R., Hertzberg, V. & Walker, M. 1989. Measurements of acute cerebral infarction: lesion size by computed tomography. *Stroke; a journal of cerebral circulation,* 20**,** 871-875.

Bruce H. Dobkin, M. D. 1989. Focused Stroke Rehabilitation Programs Do Not Improve Outcome. *Arch neurol,* 46**,** 701-703.

Buma, F., Kwakkel, G. & Ramsey, N. 2013. Understanding upper limb recovery after stroke. *Restorative neurology and neuroscience,* 31**,** 707-722.

Bustamante, A., Sobrino, T., Giralt, D., García-Berrocoso, T., Llombart, V., Ugarriza, I., Espadaler, M., Rodríguez, N., Sudlow, C., Castellanos, M., Smith, C. J., Rodríguez-Yánez, M., Waje-Andreassen, U., Tanne, D., Oto, J., Barber, M., Worthmann, H., Wartenberg, K. E., Becker, K. J., Chakraborty, B., Oh, S.-H., Whiteley, W. N., Castillo, J. & Montaner, J. 2014. Prognostic value of blood interleukin-6 in the prediction of functional outcome after stroke: a systematic review and meta-analysis. *Journal Of Neuroimmunology,* 274**,** 215-224.

Chen, G., Jaradat, S. & Banerjee, N. 2002. Evaluation and comparison of clustering algorithms in analyzing ES cell gene expression data. *Statistica ...*, 1-33.

Chen, S.-Y. & Winstein, C. J. 2009. A systematic review of voluntary arm recovery in hemiparetic stroke: critical predictors for meaningful outcomes using the international classification of functioning, disability, and health. *Journal of neurologic physical therapy : JNPT,* 33, 2-13.

Church, C., Price, C., Pandyan, A. D., Huntley, S., Curless, R. & Rodgers, H. 2006. Randomized controlled trial to evaluate the effect of surface neuromuscular electrical stimulation to the shoulder after acute stroke. *Stroke,* 37, 2995-3001.

Cieza, A., Hilfiker, R., Chatterji, S., Kostanjsek, N., Üstün, B. T. & Stucki, G. 2009. The International Classification of Functioning, Disability, and Health could be used to measure functioning. *Journal of clinical epidemiology,* 62, 899-911.

Cioncoloni, D., Martini, G., Piu, P., Taddei, S., Acampa, M., Guideri, F., Tassi, R. & Mazzocchio, R. 2013. Predictors of long-term recovery in complex activities of daily living before discharge from the stroke unit. *NeuroRehabilitation,* 33, 217-223.

Clarke, B., Fokoue, E. & Zhang, H. H. 2009. *Principles and theory for data mining and machine learning*, Springer Science & Business Media.

Collin, C. & Wade, D. 1990. Assessing motor impairment after stroke: a pilot reliability study. *Journal of Neurology, Neurosurgery & Psychiatry,* 53, 576-579.

Collins, G. S., De Groot, J. A., Dutton, S., Omar, O., Shanyinde, M., Tajar, A., Voysey, M., Wharton, R., Yu, L.-M. & Moons, K. G. 2014. External validation of multivariable prediction models: a systematic review of methodological conduct and reporting. *BMC medical research methodology,* 14, 40.

Counsell, C., Dennis, M., Mcdowall, M. & Warlow, C. 2002. Predicting outcome after acute and subacute stroke: development and validation of new prognostic models. *Stroke,* 33, 1041-1047.

Coupar, F., Pollock, A., Rowe, P., Weir, C. & Langhorne, P. 2012. Predictors of upper limb recovery after stroke: a systematic review and meta-analysis. *Clinical rehabilitation,* 26, 291-313.

Dacosta-Aguayo, R., Graña, M., Savio, A., Fernández-Andújar, M., Millán, M., López-Cancio, E., Cáceres, C., Bargalló, N., Garrido, C., Barrios, M., Clemente, I. C., Hernández, M., Munuera, J., Dávalos, A., Auer, T. & Mataró, M. 2014. Prognostic value of changes in resting-state functional connectivity patterns in cognitive recovery after stroke: A 3T fMRI pilot study. *Human Brain Mapping,* 35, 3819-3831.

David W. Hosmer , J. 2013. *Applied Logistic Regression*.

Dormann, C. F., Elith, J., Bacher, S., Buchmann, C., Carl, G., Carré, G., Marquéz, J. R. G., Gruber, B., Lafourcade, B. & Leitão, P. J. 2013. Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. *Ecography,* 36, 27-46.

Efron, B. & Tibshirani, R. J. 1994. *An introduction to the bootstrap*, CRC press.

Eghidemwivbie, N. T. & Schneeweis, V. A. 2010. Early prediction of functional outcome by physiotherapists in post stroke patients A prospective cohort study.

Enderby, P., Pandyan, A., Bowen, A., Hearnden, D., Ashburn, A., Conroy, P., Logan, P., Thompson, C. & Winter, J. 2017. Accessing rehabilitation after stroke–a guessing game? : Taylor & Francis.

Europe, S. a. F. 2017. *The Burden Of Stroke In Europe – Challenges For Policy Makers* [Online]. London: Stroke Alliance for Europe. Available: https://www.stroke.org.uk/sites/default/files/the_burden_of_stroke_in_europe_-_challenges_for_policy_makers.pdf [Accessed 18/07/2018].

Fan, Y. & Tang, C. Y. 2013. Tuning parameter selection in high dimensional penalized likelihood. *Journal of the Royal Statistical Society: Series B (Statistical Methodology),* 75, 531-552.

Fawcett, T. 2006. An introduction to ROC analysis. *Pattern recognition letters,* 27, 861-874.

Feys, H., De Weerdt, W., Nuyens, G., Van De Winckel, A., Selz, B. & Kiekens, C. 2000a. Predicting motor recovery of the upper limb after stroke rehabilitation: value of a clinical examination. *Physiotherapy research international : the journal for researchers and clinicians in physical therapy,* 5**,** 1-18.

Feys, H., Hetebrij, J., Wilms, G., Dom, R. & De Weerdt, W. 2000b. Predicting arm recovery following stroke: value of site of lesion. *Acta neurologica Scandinavica,* 102**,** 371-377.

Feys, H. M., De Weerdt, W. J., Selz, B. E., Cox Steck, G. A., Spichiger, R., Vereeck, L. E., Putman, K. D. & Van Hoydonck, G. A. 1998. Effect of a Therapeutic Intervention for the Hemiplegic Upper Limb in the Acute Phase After Stroke : A Single-Blind, Randomized, Controlled Multicenter Trial. *Stroke,* 29**,** 785-792.

Fleuren, J. F., Buurke, J. H. & Geurts, A. C. 2018. Clinical Management of Spasticity and Contractures in Stroke. *Neurological Rehabilitation.* CRC Press.

Frank E Harrell , J. 2001. <Regression Modeling Strategies With Applications to Linear Models, Logistic Regression, and Survival Analysis.pdf>.

Friedman, J. H., Hastie, T. & Tibshirani, R. glmnet: lasso and elastic-net regularized generalized linear models, 2010b. *URL http://CRAN.* R-project. org/package= glmnet. *R package version***,** 1.1-5.

Fritz, S. L., Blanton, S., Uswatte, G., Taub, E. & Wolf, S. L. 2009. Minimal detectable change scores for the Wolf Motor Function Test. *Neurorehabilitation and neural repair,* 23**,** 662-667.

Fu, T. S. T., Wu, C. Y., Lin, K. C., Hsieh, C. J., Liu, J. S., Wang, T. N. & Ou-Yang, P. 2012. Psychometric comparison of the shortened Fugl-Meyer Assessment and the streamlined Wolf Motor Function Test in stroke rehabilitation. *Clinical rehabilitation,* 26**,** 1043-1047.

Fugl-Meyer, A. R., Jääskö, L., Leyman, I., Olsson, S. & Steglind, S. 1975. The post-stroke hemiplegic patient. 1. a method for evaluation of physical performance. *Scandinavian Journal Of Rehabilitation Medicine,* 7**,** 13-31.

Gao, X., Pu, D. Q., Wu, Y. & Xu, H. 2012. Tuning parameter selection for penalized likelihood estimation of Gaussian graphical model. *Statistica Sinica***,** 1123-1146.

Gebruers, N., Truijen, S., Engelborghs, S. & De Deyn, P. P. 2014. Prediction of upper limb recovery, general disability, and rehabilitation status by activity measurements assessed by accelerometers or the Fugl-Meyer score in acute stroke. *American journal of physical medicine & rehabilitation / Association of Academic Physiatrists,* 93**,** 245-252.

Gert Kwakkela, B., ∗ and Boudewijn Kollenc 2007. Predicting improvement in the upper paretic limb after stroke: A longitudinal prospective study. *Restorative neurology and neuroscience*.

Giardini, A., Ferrari, P., Majani, G., Negri, E. M., Rossi, S., Magnani, C. & Preti, P. 2010. International Classification of Functioning Disability and Health (ICF) e Qualit{à} della Vita nel paziente oncologico in fase avanzata di malattia. *G Ital Med Lav Ergon,* 32**,** B29--36.

Goeman, J., Meijer, R. & Chaturvedi, N. 2012. L1 and L2 penalized regression models. *Cran. R-Project. or***,** 1-20.

Gotay, C. C. & Wilson, M. 1998. Use of quality-of-life outcome assessments in current cancer clinical trials. *Evaluation & the health professions,* 21**,** 157-178.

Green, J. B. 2003. Brain reorganization after stroke. *Topics in stroke rehabilitation,* 10**,** 1-20.

Guo, P., Zeng, F., Hu, X., Zhang, D., Zhu, S., Deng, Y. & Hao, Y. 2015. Improved variable selection algorithm using a LASSO-type penalty, with an application to assessing Hepatitis B infection relevant factors in community residents. *Plos One,* 10**,** e0134151.

Harrell, F. E. & Lee, K. L. 1984. Regression Modelling Strategies For Improved Prognostic Prediction. 3**,** 143-152.

Harrell, F. E., Lee, K. L. & Mark, D. B. 1996. Multivariable prognostic models: Issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in Medicine,* 15**,** 361-387.

Harrell Jr, F. E. 2015. *Regression modeling strategies: with applications to linear models, logistic and ordinal regression, and survival analysis*, Springer.

Hastie, T., Tibshirani, R. & Wainwright, M. 2015. *Statistical learning with sparsity: the lasso and generalizations*, CRC press.

Hendricks, H. T., Hageman, G. & Van Limbeek, J. 1997. Prediction of recovery from upper extremity paralysis after stroke by measuring evoked potentials. *Scandinavian Journal Of Rehabilitation Medicine,* 29**,** 155-159.

Hennerici, M. G. 2004. The unstable plaque. *Cerebrovascular Diseases,* 17**,** 17-22.

Holmberg, L. & Vickers, A. 2013. Evaluation of prediction models for decision-making: beyond calibration and discrimination. *PLoS medicine,* 10**,** e1001491.

Houwink, A., Nijland, R. H., Geurts, A. C. & Kwakkel, G. 2013. Functional recovery of the paretic upper limb after stroke: who regains hand capacity? *Archives of Physical Medicine and Rehabilitation,* 94**,** 839-844.

Hsieh, C. L., Hsueh, I. P., Chiang, F. M. & Lin, P. H. 1998. Inter-rater reliability and validity of the Action Research arm test in stroke patients. *Age and Ageing,* 27**,** 107-114.

Hsieh, Y.-W., Wu, C.-Y., Lin, K.-C., Chang, Y.-F., Chen, C.-L. & Liu, J.-S. 2009. Responsiveness and validity of three outcome measures of motor function after stroke rehabilitation. *Stroke; a journal of cerebral circulation,* 40**,** 1386-1391.

Jack, S. S. 1981. Current estimates from the National Health Interview Survey: United States 1980. *Vital and Health Statistics Series 10: Data From the National Health Survey***,** 1-83.

James, G., Witten, D., Hastie, T. & Tibshirani, R. 2013. *An introduction to statistical learning*, Springer.

Kassambara, A. 2017. *Practical guide to cluster analysis in R: Unsupervised machine learning*, STHDA.

Katrak, P., Bowring, G., Conroy, P., Chilvers, M., Poulos, R. & Mcneil, D. 1998. Predicting upper limb recovery after stroke: The place of early shoulder and hand movement. *Archives of Physical Medicine and Rehabilitation,* 79**,** 758-761.

Keidan, I., Shahar, E., Barzilay, Z., Passwell, J. & Brand, N. 1994. Predictors of outcome of stroke in infants and children based on clinical data and radiologic correlates. *Acta Paediatrica (Oslo, Norway: 1992),* 83**,** 762-765.

Kkel, K. G., Kollen, B., Lindeman, E., Kwakkel, G., Kollen, B. & Lindeman, E. 2004. Understanding the pattern of functional recovery after stroke: facts and theories. *Restorative neurology and neuroscience,* 22**,** 281-299.

Kong, K.-H. & Lee, J. 2013. Temporal recovery and predictors of upper limb dexterity in the first year of stroke: a prospective study of patients admitted to a rehabilitation centre. *NeuroRehabilitation,* 32**,** 345-350.

Koyama, T., Matsumoto, K., Okuno, T. & Domen, K. 2005. A new method for predicting functional recovery of stroke patients with hemiplegia: logarithmic modelling. *Clinical rehabilitation,* 19**,** 779-789.

Kumar, P., Yadav, A. K., Misra, S., Kumar, A., Chakravarty, K. & Prasad, K. 2016. Prediction of upper extremity motor recovery after subacute intracerebral hemorrhage through diffusion tensor imaging: a systematic review and meta-analysis. *Neuroradiology,* 18**,** 50-59.

Kundu, S., Aulchenko, Y. S., Van Duijn, C. M. & Janssens, A. C. J. 2011. PredictABEL: an R package for the assessment of risk prediction models. *European journal of epidemiology,* 26**,** 261.

Kutner H.Micheal, N. J. C. N. J. W. L. 2005. *applied linear statistical models,* new york, mcGraw-hill.

Kwah, L. K. & Diong, J. 2014. National Institutes of Health Stroke Scale (NIHSS). *Journal of Physiotherapy,* 60**,** 61-61.

Kwah, L. K., Harvey, L. A., Diong, J. & Herbert, R. D. 2013. Models containing age and NIHSS predict recovery of ambulation and upper limb function six months after stroke: An observational study. *Journal of Physiotherapy,* 59**,** 189-197.

Kwah, L. K. & Herbert, R. D. 2016. Prediction of Walking and Arm Recovery after Stroke: A Critical Review. *Brain sciences,* 6**,** 53.

Kwakkel, G. 2009. Intensity of practice after stroke: More is better. *Schweizer Archiv fur Neurologie und Psychiatrie,* 160**,** 295-298.

Kwakkel, G. & Kollen, B. 2007. Predicting improvement in the upper paretic limb after stroke: A longitudinal prospective study. *Restorative Neurology & Neuroscience,* 25**,** 453-460.

Kwakkel, G., Kollen, B. & Twisk, J. 2006. Impact of time on improvement of outcome after stroke. *Stroke,* 37**,** 2348-2353.

Kwakkel, G. & Kollen, B. J. 2013. Predicting activities after stroke: What is clinically relevant? *International Journal of Stroke,* 8**,** 25-32.

Kwakkel, G., Kollen, B. J. & Krakauer, J. W. 2014. Predicting activities after stroke. *Oxford Textbook of Neurorehabilitation.* Oxford University Press.

Kwakkel, G., Kollen, B. J., Van Der Grond, J. V. & Prevo, A. J. H. 2003. Probability of regaining dexterity in the flaccid upper limb: Impact of severity of paresis and time since onset in acute stroke. *Stroke,* 34**,** 2181-2186.

Kwakkel, G., Van Dijk, G. M. & Wagenaar, R. C. 2000. Accuracy of physical and occupational therapists' early predictions of recovery after severe middle cerebral artery stroke. *Clinical rehabilitation,* 14**,** 28-41.

Kwakkel, G., Wagenaar, R. C., Kollen, B. J. & Lankhorst, G. J. 1996. Predicting disability in stroke--a critical review of the literature. *Age and Ageing,* 25**,** 479-489.

Leathley, M. J., Gregson, J., Moore, A., Smith, T., Sharma, A. & Watkins, C. 2004. Predicting spasticity after stroke in those surviving to 12 months. *Clinical rehabilitation,* 18**,** 438-443.

Levin, M. F., Kleim, J. A. & Wolf, S. L. 2009. What Do Motor “Recovery” and “Compensation” Mean in Patients Following Stroke? 23**,** 313-319.

Lin, K.-C., Huang, Y.-H., Hsieh, Y.-W. & Wu, C.-Y. 2009. Potential predictors of motor and functional outcomes after distributed constraint-induced therapy for patients with stroke. *Neurorehabilitation and neural repair,* 23**,** 336-342.

Lindsay, C., Simpson, J., Ispoglou, S., Sturman, S. G. & Pandyan, A. D. 2014. The early use of botulinum toxin in post-stroke spasticity: study protocol for a randomised controlled trial. *Trials,* 15**,** 12.

Loewen, S. C. & Anderson, B. A. 1990. Predictors of stroke outcome using objective measurement scales. *Stroke; a journal of cerebral circulation,* 21**,** 78-81.

Mathiowetz, V., Volland, G., Kashman, N. & Weber, K. 1985. Adult norms for the Box and Block Test of manual dexterity. *American Journal of Occupational Therapy,* 39**,** 386-391.

Mathiowetz, V. & Weber, K. 1985. Adult N o rills for the Box and Block. 39.

Mcginn, T. G., Guyatt, G. H., Wyer, P. C., Naylor, C. D., Stiell, I. G., Richardson, W. S. & Grp, E.-B. M. W. 2000. Users' guides to the medical literature - XXII: How to use articles about clinical decision rules. *Jama-Journal of the American Medical Association,* 284**,** 79-84.

Mead, G., Lewis, S., Wardlaw, J., Dennis, M. & Warlow, C. 2000. How well does the Oxfordshire Community Stroke Project classification predict the site and size of the infarct on brain imaging? *Journal of Neurology, Neurosurgery & Psychiatry,* 68**,** 558-562.

Meier, L., Geer, S. V. D. & Bühlmann, P. 2008. The group lasso for logistic regression. *J.R. Statist. Soc. B,* 70**,** 53-71.

Meyer, S., Karttunen, A. H., Thijs, V., Feys, H. & Verheyden, G. 2014. How Do Somatosensory Deficits in the Arm and Hand Relate to Upper Limb Impairment, Activity, and Participation Problems After Stroke? A Systematic Review. *Physical Therapy*.

Mille, A. 2002. *Subset Selection in Regression*.

Ming Yuan, A. L. 2006. Model selection and estimation in regression with grouped variables.

Mortimer, J. & Green, M. 2015. Briefing: the health and care of older people in England 2015. *Age UK, London.* http://www.ageuk.org.uk/professional-resources-home/research/reports/care-and-support/the-health-and-care-of-older-people-in-england-2015/*(accessed 22 September 2016)*.

Moura, R. D. C. D. R., Fukujima, M. M., Aguiar, A. S., Fontes, S. V., Dauar, R. F. B. & Prado, G. F. D. 2009. Predictive factors for spasticity among ischemic stroke patients. *Arquivos de neuro-psiquiatria,* 67**,** 1029-1036.

Murray, C. J. L., Vos, T., Lozano, R., Naghavi, M., Flaxman, A. D., Michaud, C., Ezzati, M., Shibuya, K., Salomon, J. A., Abdalla, S., Aboyans, V., Abraham, J., Ackerman, I., Aggarwal, R., Ahn, S. Y., Ali, M. K., Alvarado, M., Anderson, H. R., Anderson, L. M., Andrews, K. G., Atkinson, C., Baddour, L. M., Bahalim, A. N., Barker-Collo, S., Barrero, L. H., Bartels, D. H., Basáñez, M. G., Baxter, A., Bell, M. L., Benjamin, E. J., Bennett, D., Bernabé, E., Bhalla, K., Bhandari, B., Bikbov, B., Abdulhak, A. B., Birbeck, G., Black, J. A., Blencowe, H., Blore, J. D., Blyth, F., Bolliger, I., Bonaventure, A., Boufous, S., Bourne, R., Boussinesq, M., Braithwaite, T., Brayne, C., Bridgett, L., Brooker, S., Brooks, P., Brugha, T. S., Bryan-Hancock, C., Bucello, C., Buchbinder, R., Buckle, G., Budke, C. M., Burch, M., Burney, P., Burstein, R., Calabria, B., Campbell, B., Canter, C. E., Carabin, H., Carapetis, J., Carmona, L., Cella, C., Charlson, F., Chen, H., Cheng, A. T. A., Chou, D., Chugh, S. S., Coffeng, L. E., Colan, S. D., Colquhoun, S., Colson, K. E., Condon, J., Connor, M. D., Cooper, L. T., Corriere, M., Cortinovis, M., De Vaccaro, K. C., Couser, W., Cowie, B. C., Criqui, M. H., Cross, M., Dabhadkar, K. C., Dahiya, M., Dahodwala, N., Damsere-Derry, J., Danaei, G., Davis, A., De Leo, D., Degenhardt, L., Dellavalle, R., Delossantos, A., Denenberg, J., Derrett, S., Des Jarlais, D. C., Dharmaratne, S. D., et al. 2012. Disability-adjusted life years (DALYs) for 291 diseases and injuries in 21 regions, 1990-2010: A systematic analysis for the Global Burden of Disease Study 2010. *The Lancet,* 380**,** 2197-2223.

Nijland, R., Van Wegen, E., Verbunt, J., Van Wijk, R., Van Kordelaar, J. & Kwakkel, G. 2010a. A comparison of two validated tests for upper limb function after stroke: The Wolf Motor Function Test and the Action Research Arm Test. *Journal of Rehabilitation Medicine,* 42**,** 694-696.

Nijland, R. H. M., Van Wegen, E. E. H., Harmeling-Van Der Wel, B. C. & Kwakkel, G. 2010b. Presence of finger extension and shoulder abduction within 72 hours after stroke predicts functional recovery: early prediction of functional outcome after stroke: the EPOS cohort study. *Stroke; a journal of cerebral circulation,* 41**,** 745-750.

Nijland, R. H. M., Van Wegen, E. E. H., Harmeling-Van Der Wel, B. C. & Kwakkel, G. 2010c. Presence of finger extension and shoulder abduction within 72 hours after stroke predicts functional recovery: Early prediction of functional outcome after stroke: The EPOS cohort study. *Stroke,* 41**,** 745-750.

Nijland, R. H. M., Van Wegen, E. E. H., Harmeling-Van Der Wel, B. C. & Kwakkel, G. 2013. Accuracy of physical therapists' early predictions of upper-limb function in hospital stroke units: the EPOS Study. *Physical Therapy,* 93**,** 460-469.

Nordin, A., Murphy, M. A. & Danielsson, A. 2014. Intra-rater and inter-rater reliability at the item level of the Action Research Arm Test for patients with stroke. *Journal of Rehabilitation Medicine***,** 738-745.

Ohura, T., Hase, K., Nakajima, Y. & Nakayama, T. 2017. Validity and reliability of a performance evaluation tool based on the modified Barthel Index for stroke patients. *BMC medical research methodology,* 17**,** 131.

Opheim, A., Danielsson, A., Murphy, M. A., Persson, H. C. & Sunnerhagen, K. S. 2015. Early prediction of long-term upper limb spasticity after stroke Part of the SALGOT study. *Neurology***,** 10.1212/WNL. 0000000000001908.

Pan, J. & Shang, J. 2017. Adaptive LASSO for linear mixed model selection via profile log-likelihood. *Communications in Statistics - Theory and Methods***,** 1-19.

Pandyan, A. D., Hermens, H. J. & Conway, B. A. 2018. *Neurological Rehabilitation: Spasticity and Contractures in Clinical Practice and Research*, CRC Press.

Pinter, M. M. & Brainin, M. 2012. Rehabilitation after stroke in older people. *Maturitas,* 71**,** 104-108.

Platz, T., Pinkowski, C., Van Wijck, F., Kim, I.-H., Di Bella, P. & Johnson, G. 2005. Reliability and validity of arm function assessment with standardized guidelines for the Fugl-Meyer

Test, Action Research Arm Test and Box and Block Test: a multicentre study. *Clinical rehabilitation,* 19**,** 404-411.

Pollock, A., Farmer, S. E., Brady, M. C., Langhorne, P., Mead, G. E., Mehrholz, J. & Van Wijck, F. 2015. Cochrane Overview: Figure. *Stroke,* 46**,** e57-e58.

Power, M., Bullinger, M. & Harper, A. 1999. The World Health Organization WHOQOL-100: Tests of the universality of quality of life in 15 different cultural groups worldwide. *Health psychology,* 18**,** 495.

Rand, D. & Eng, J. J. 2015. Predicting daily use of the affected upper extremity 1 year after stroke. *Journal of Stroke and Cerebrovascular Diseases,* 24**,** 274-283.

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C. & Müller, M. 2011. pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC bioinformatics,* 12**,** 77.

Rousseeuw, P. J. 1987. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics,* 20**,** 53-65.

Royston, P., Moons, K. G., Altman, D. G. & Vergouwe, Y. 2009. Prognosis and prognostic research: developing a prognostic model. *Bmj,* 338**,** b604.

Ryan, T. P. 2008. *Modern regression methods*, John Wiley & Sons.

Sacco, R. L., Kasner, S. E., Broderick, J. P., Caplan, L. R., Connors, J. J., Culebras, A., Elkind, M. S. V., George, M. G., Hamdan, A. D., Higashida, R. T., Hoh, B. L., Janis, L. S., Kase, C. S., Kleindorfer, D. O., Lee, J. M., Moseley, M. E., Peterson, E. D., Turan, T. N., Valderrama, A. L. & Vinters, H. V. 2013. An updated definition of stroke for the 21st century: A statement for healthcare professionals from the American heart association/American stroke association. *Stroke,* 44**,** 2064-2089.

Sayad, D. S. 2010-2019. *An Introduction to Data Science* [Online]. Available: https://www.saedsayad.com/clustering_hierarchical.htm [Accessed].

Schiemanck, S. K., Kwakkel, G., Post, M. W. M., Kappelle, L. J. & Prevo, A. J. H. 2006. Predicting long-term independency in activities of daily living after middle cerebral artery stroke: Does information from MRI have added predictive value compared with clinical information? *Stroke,* 37**,** 1050-1054.

Schwarz, G. 1978. Estimating the dimension of a model. *The Annals of Statistics,* 6**,** 461-464.

Simpson, L. A. & Eng, J. J. 2012. Functional Recovery Following Stroke: Capturing Changes in Upper-Extremity Function. *Neurorehabilitation and neural repair*.

Sing, T., Beerenwinkel, N. & Lengauer, T. 2004. Learning mixtures of localized rules by maximizing the area under the ROC curve. *ROCAI,* 4**,** 96-98.

Smania, N., Paolucci, S., Tinazzi, M., Borghero, A., Manganotti, P., Fiaschi, A., Moretto, G., Bovi, P. & Gambarin, M. 2007. Active finger extension: a simple movement predicting recovery of arm function in patients with acute stroke. *Stroke; a journal of cerebral circulation,* 38**,** 1088-1090.

Sterne, J. A., White, I. R., Carlin, J. B., Spratt, M., Royston, P., Kenward, M. G., Wood, A. M. & Carpenter, J. R. 2009. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *Bmj,* 338**,** b2393.

Steyerberg, E. W. 2009. *Clinical Prediction Models A Practical Approach to Development, Validation, and Updating*.

Steyerberg, E. W. & Vergouwe, Y. 2014. Towards better clinical prediction models: seven steps for development and an ABCD for validation. *European heart journal,* 35**,** 1925-1931.

Steyerberg, E. W., Vickers, A. J., Cook, N. R., Gerds, T., Gonen, M., Obuchowski, N., Pencina, M. J. & Kattan, M. W. 2010a. Assessing the performance of prediction models: a framework for some traditional and novel measures. *Epidemiology (Cambridge, Mass.),* 21**,** 128.

Steyerberg, E. W., Vickers, A. J., Cook, N. R., Gerds, T., Gonen, M., Obuchowski, N., Pencina, M. J. & Kattan, M. W. 2010b. Assessing the performance of prediction models: a framework for traditional and novel measures. *Epidemiology,* 21**,** 128-38.

Stinear, C. 2010. Prediction of recovery of motor function after stroke. *The Lancet. Neurology,* 9**,** 1228-1232.

Stinear, C. M., Barber, P. A., Petoe, M., Anwar, S. & Byblow, W. D. 2012. The PREP algorithm predicts potential for upper limb recovery after stroke. *Brain: A Journal Of Neurology,* 135**,** 2527-2535.

Stroke Association 2018. State of the nation.

Sun, W., Wang, J. & Fang, Y. 2013. Consistent selection of tuning parameters via variable selection stability. *The Journal of Machine Learning Research,* 14**,** 3419-3440.

Sun, X. & Xu, W. 2014. Fast implementation of DeLong's algorithm for comparing the areas under correlated receiver operating characteristic curves. *IEEE Signal Processing Letters,* 21**,** 1389-1393.

Sunnerhagen, K. S. 2016. Predictors of spasticity after stroke. *Current physical medicine and rehabilitation reports,* 4**,** 182-185.

Suzuki, M., Omori, M., Hatakeyama, M., Yamada, S., Matsushita, K. & Iijima, S. 2006. Predicting Recovery of Upper-Body Dressing Ability After Stroke. *Archives of Physical Medicine and Rehabilitation,* 87**,** 1496-1502.

Suzuki, M., Omori, Y., Sugimura, S., Miyamoto, M., Sugimura, Y., Kirimoto, H. & Yamada, S. 2011. Predicting recovery of bilateral upper extremity muscle strength after stroke. *Journal of Rehabilitation Medicine,* 43**,** 935-943.

Suzuki, M., Sugimura, Y., Yamada, S., Omori, Y., Miyamoto, M. & Yamamoto, J. I. 2013. Predicting Recovery of Cognitive Function Soon after Stroke: Differential Modeling of Logarithmic and Linear Regression. *PLoS ONE,* 8**,** 26-28.

Taub, E., Morris, D. M., Crago, J., King, D. K., Bowman, M., Bryson, C., Bishop, S., Pearson, S. & Shaw, S. E. 2011. Wolf Motor Function Test (WMFT) Manual written by. *Therapy***,** 1-31.

Tibshirani, R. 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)***,** 267-288.

Tibshirani, R., Walther, G. & Hastie, T. 2001. Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology),* 63**,** 411-423.

Tilling, K., Sterne, J. A., Rudd, A. G., Glass, T. A., Wityk, R. J. & Wolfe, C. D. 2001a. A new method for predicting recovery after stroke. *Stroke; a journal of cerebral circulation,* 32**,** 2867-2873.

Tilling, K., Sterne, J. A. & Wolfe, C. D. 2001b. Multilevel growth curve models with covariate effects: application to recovery after stroke. *Statistics in Medicine,* 20**,** 685-704.

Trevor Hastie, R. T., Jerome Friedman 2015. *Statistical Learning with Sparsity The Lasso and Generalizations*.

Vach, W. 2013. Calibration of clinical prediction rules does not just assess bias. *Journal of clinical epidemiology,* 66**,** 1296-1301.

Van Calster, B., Nieboer, D., Vergouwe, Y., Pencina, M. J. & Steyerberg, E. W. 2015. A calibration hierarchy for risk models: strong calibration occurs only in utopia.

Van Calster, B., Wynants, L., Verbeek, J. F. M., Verbakel, J. Y., Christodoulou, E., Vickers, A. J., Roobol, M. J. & Steyerberg, E. W. 2018. Reporting and Interpreting Decision Curve Analysis: A Guide for Investigators. *Eur Urol,* 74**,** 796-804.

Van Der Lee, J. H., Beckerman, H., Lankhorst, G. J. & Bouter, L. M. 2001. The responsiveness of the Action Research Arm test and the Fugl-Meyer Assessment scale in chronic stroke patients. *Journal of Rehabilitation Medicine,* 33**,** 110-113.

Van Der Lee, J. H., Beckerman, H., Lankhorst, G. J., Bouter, L. M., Nordin, A., Murphy, M. A., Danielsson, A., Nijland, R., Van Wegen, E., Verbunt, J., Van Wijk, R., Van Kordelaar, J., Kwakkel, G., Markowitsch, H. J., Calabrese, P., Bielefeld, D., Koh, C.-L., Hsueh, I. P., Wang, W.-C., Sheu, C.-F., Yu, T.-Y., Wang, C.-H., Hsieh, C.-L., Jung, T.-D., Kim, J.-Y., Seo, J.-H., Jin, S.-U., Lee, H. J., Lee, S.-H., Lee, Y.-S., Chang, Y., Iosa, M., Morone, G., Fusco, A., Paolucci, S., Control, I., Ashford, S., Slade, M., Malaprade, F. & Turner-Stokes, L. 2010. A comparison of two validated tests for upper limb function after stroke: The wolf motor function test and the action research arm test. *Journal of Rehabilitation Medicine,* 42**,** 694-696.

Veerbeek, J. M., Kwakkel, G., Van Wegen, E. E. H., Ket, J. C. F. & Heymans, M. W. 2011. Early prediction of outcome of activities of daily living after stroke: A systematic review. *Stroke,* 42**,** 1482-1488.

Veerbeek, J. M., Van Wegen, E., Van Peppen, R., Van Der Wees, P. J., Hendriks, E., Rietberg, M. & Kwakkel, G. 2014. What is the evidence for physical therapy poststroke? A systematic review and meta-analysis. *Plos One,* 9**,** e87987.

Vickers, A. J. & Elkin, E. B. 2006. Decision curve analysis: a novel method for evaluating prediction models. *Medical Decision Making,* 26**,** 565-574.

Wang, X. & Fan, J. 2014. Variable selection for multivariate generalized linear models. *Journal of Applied Statistics,* 41**,** 393-406.

Weimar, C., Ziegler, A., König, I. R. & Diener, H.-C. 2002. Predicting functional outcome and survival after acute ischemic stroke. *Journal of Neurology,* 249**,** 888-895.

Who 2001. *International classification of functioning, disability and health: ICF*, Geneva: World Health Organization.

Woldag, H., Gerhold, L. L., De Groot, M., Wohlfart, K., Wagner, A. & Hummelsheim, H. 2006. Early prediction of functional outcome after stroke. *Brain Injury,* 20**,** 1047-1052.

Wolf, S. L., Lecraw, D. E., Barton, L. A. & Jann, B. B. 1989. Forced use of hemiplegic upper extremities to reverse the effect of learned nonuse among chronic stroke and head-injured patients. *Experimental Neurology,* 104**,** 125-132.

Wolfe, C. D. 2000. The impact of stroke. *British medical bulletin,* 56**,** 275-286.

Yagura, H., Miyai, I., Seike, Y., Suzuki, T. & Yanagihara, T. 2003. Benefit of inpatient multidisciplinary rehabilitation up to 1 year after stroke11No commercial party having a direct financial interest in the results of the research supporting this article has or will confer a benefit upon the author(s) or upon any orga. *Archives of Physical Medicine and Rehabilitation,* 84**,** 1687-1691.

Yozbatiran, N., Der-Yeghiaian, L. & Cramer, S. C. 2008. A standardized approach to performing the action research arm test. *Neurorehabilitation and neural repair,* 22**,** 78-90.

Yuan, M. & Lin, Y. 2006. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology),* 68**,** 49-67.

Zhang, N., Liu, G., Zhang, G., Fang, J., Wang, Y., Zhao, X., Pan, Y., Guo, L., Wang, Y. & China National Stroke Registry, I. 2013. External validation of the iScore for predicting ischemic stroke mortality in patients in China. *Stroke,* 44**,** 1924-9.

Zhang, S., Lu, Y., Zhang, L., Cai, B. & Qiu, K. 2015. Variable Selection in Logistic Regression Model. *Chinese Journal of Electronics,* 24**,** 813-817.

Zhang, Y., Li, R. & Tsai, C.-L. 2010. Regularization parameter selections via generalized information criterion. *Journal of the American statistical association,* 105**,** 312-323.

Zhang, Z., Rousson, V., Lee, W.-C., Ferdynus, C., Chen, M., Qian, X. & Guo, Y. 2018. Decision curve analysis: a technical note. *Annals of translational medicine,* 6.

Zou, H. 2006. The adaptive lasso and its oracle properties. *Journal of the American statistical association,* 101**,** 1418-1429.

10 Appendices

10.1 Appendix A

Publications

Three Abstract have published in the international journal of stroke, which is owned by the SAGE journals. The work within this publication has been actively discussed at the UK forum conferences 2016 and 2017.

“Prediction of upper limb function recovery post-stroke”

Al-Shallawi A1, Blana D1 and Pandyan A1,2

Introduction: Stroke can lead to a loss of arm function and this can severely affect a person’s life. Predicting recovery post-stroke can be very beneficial to stroke patients and medical professionals. Specificity and sensitivity of current models are not good enough predicting accurately and they are not adequate for implementation into routine clinical practice. The aim of this study is to explore if methods of clustering can help improve sensitivity and specificity of prediction models.

Method: Retrospective modelling on a secondary anonymised data set was undertaken. The dependent variable was arm function measured using Action Research Arm Test (ARAT). The independent variables were NIHSS score, Frenchay Arm Test (FAT), Motricity Index (MI) and age. A logistic regression model was developed for the entire sample set and specificity and sensitivity were quantified. This process was repeated after using k-means clustering method procedure.

Results: The logistic regression model demonstrated that the NIHSS, FAT and MI before clustering analysis were able to classify probability of recovery with a sensitivity of 0.90 and specificity of 0.93 ($p < 0.0001$). The k-means clustering produced 4 homogenous clusters of patients. In 3 of the 4 groups the sensitivity and specificity were 1 ($p < 0.0001$). In the fourth cluster it was 0.97 ($p < 0.0001$)

Conclusion: Using methods of clustering may provide a better approach to modelling recovery after stroke. However more work is needed to confirm the reliability and clinical usefulness of the methods of clustering.

"Improving variable selection for modelling recovery of upper limb function post-stroke"

Al-Shallawi A[1], Blana D[1] and Pandyan A[2]

Introduction: Loss of arm function post-stroke can severely affect a person's life. Predicting recovery prospectively is difficult, particularly in patients with severe levels of initial impairments. The inherent variance associated with variable selection within the traditional methods of modelling could be a reason for this. Newer methods of modelling that use unbiased methods of selecting variables and modelling are now available (Lasso, Adaptive Lasso and group Lasso). The aim is to compare these new methods against the traditional methods.

Methods: A database of 150 stroke patients was analysed. Each patient had baseline measurements taken within a week of a stroke (giving 78 independent variables), and arm function measurements (Action Research Arm Test – ARAT) taken at 12 weeks after stroke (the dependent variable). Stepwise logistic regression, Lasso and Adaptive Lasso were used for variable selection and modelling. Results: Lasso, Adaptive lasso and group Lasso shrunk 78 predictors to 8, 6 and 11 predictors respectively with accuracy (87%, 88% and 87%), sensitivity (95%, 95% and 93%), specificity (0.67, 0.74 and 0.62) and F-measure (0.73, 0.76 and 0.73). The traditional method selected the 3 variables which were not significantly related to the clinical treatment with 85% sensitivity; 56% specificity; 0.53 F-measure.

Conclusion: It is evident from the results that the newer methods could conceivably be employed in selecting predictors to develop a prediction model of recovery upper limb post-stroke. These could improve the clinical usefulness.

“What cut-off is indicative of no upper limb function in the Action Research Arm Test?”

Al-Shallawi A1, Blana D1 and Pandyan A2

Introduction: The Action Research Arm Test (ARAT) is a clinical scale that is used for assessing the upper limb (UL) function of stroke survivors. Previous studies have reported that patients who have a total score of less than 9 can be classified as having severely limited UL function. The aim of this study was to investigate if this current ‘‘cut-off’’ is valid.

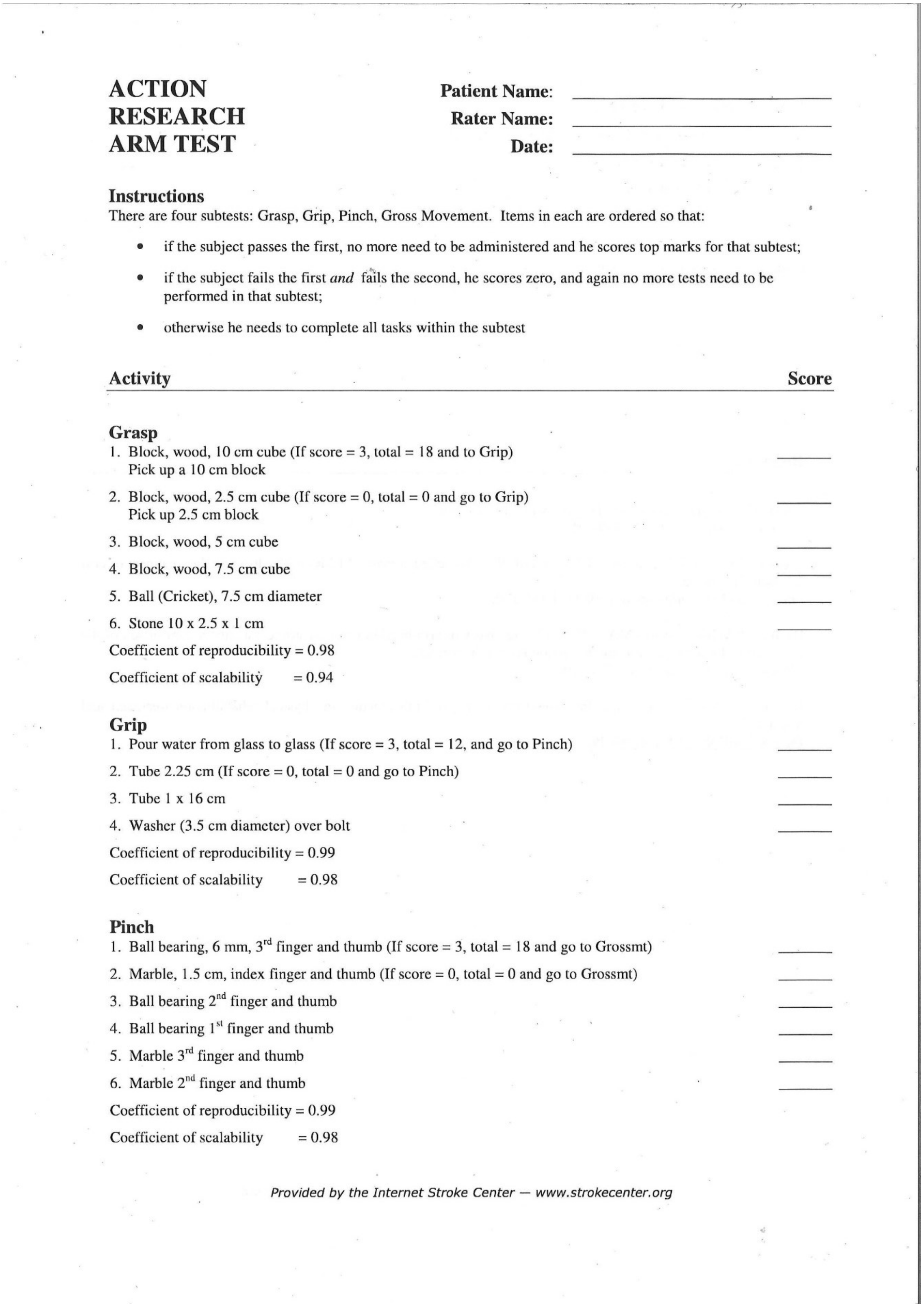
Methods: Retrospective analysis of the ARAT scores from secondary anonymised data set with 150 participants. The baseline measures, taken within 1 week of a stroke, informed the analysis.

Results: 66 (44% with 95% Confidence interval (CI) 36% to 52%) patients had an ARAT score of 0 and were removed. 18 had a score between 1 and 9 (12%; 95% CI 7.5% to 18.6%). 66 had a score > 10 (95% CI 35% to 52%). Within the subset who scored between 1 and 9; 8 (44% with 95% CI 22% to 69%) were able to carry out simulated grasping tasks, 2 (11% with 95% CI 2% to 36%) could carry out simulated grip tasks, and 1 (1.5% with 95% CI 0.3% to 3%) patients achieved a score in the simulated pinch sub-category. If the cut-off was reduced from 9 to 7, then no person could do any of the grasp, grip and pinch subtest of the ARAT.

Conclusion: The previously used cut-off point of 9 may inappropriately classify people as non-functional. The lower cut-off of 7 should be investigated further.

10.2 Appendix B

Measures formula



ACTION RESEARCH ARM TEST

Patient Name: ______________________

Rater Name: ______________________

Date: ______________________

Instructions

There are four subtests: Grasp, Grip, Pinch, Gross Movement. Items in each are ordered so that:

- if the subject passes the first, no more need to be administered and he scores top marks for that subtest;
- if the subject fails the first *and* fails the second, he scores zero, and again no more tests need to be performed in that subtest;
- otherwise he needs to complete all tasks within the subtest

Activity **Score**

Grasp

1. Block, wood, 10 cm cube (If score = 3, total = 18 and to Grip) ______
   Pick up a 10 cm block
2. Block, wood, 2.5 cm cube (If score = 0, total = 0 and go to Grip) ______
   Pick up 2.5 cm block
3. Block, wood, 5 cm cube ______
4. Block, wood, 7.5 cm cube ______
5. Ball (Cricket), 7.5 cm diameter ______
6. Stone 10 x 2.5 x 1 cm ______

Coefficient of reproducibility = 0.98

Coefficient of scalability = 0.94

Grip

1. Pour water from glass to glass (If score = 3, total = 12, and go to Pinch) ______
2. Tube 2.25 cm (If score = 0, total = 0 and go to Pinch) ______
3. Tube 1 x 16 cm ______
4. Washer (3.5 cm diameter) over bolt ______

Coefficient of reproducibility = 0.99

Coefficient of scalability = 0.98

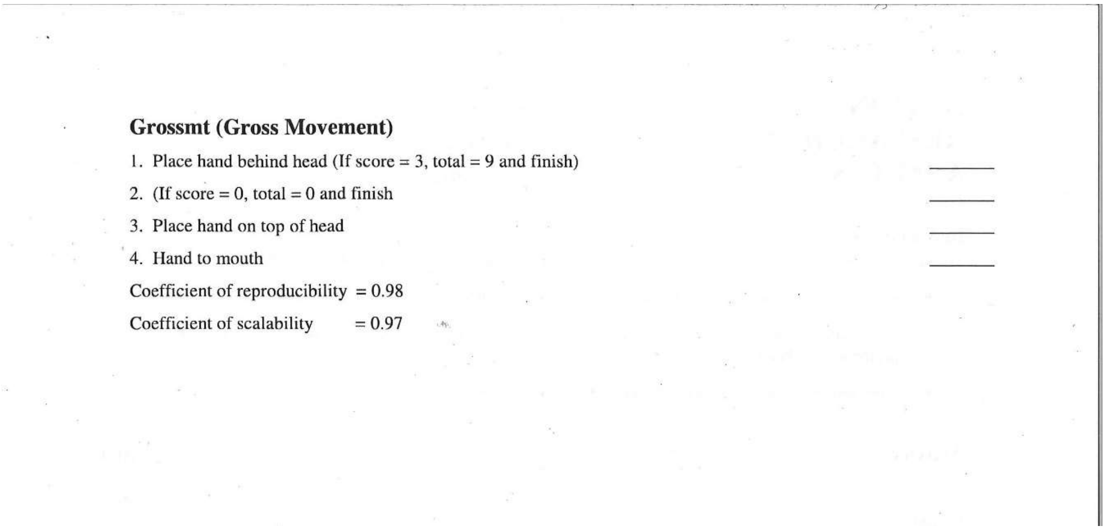
Pinch

1. Ball bearing, 6 mm, 3rd finger and thumb (If score = 3, total = 18 and go to Grossmt) ______
2. Marble, 1.5 cm, index finger and thumb (If score = 0, total = 0 and go to Grossmt) ______
3. Ball bearing 2nd finger and thumb ______
4. Ball bearing 1st finger and thumb ______
5. Marble 3rd finger and thumb ______
6. Marble 2nd finger and thumb ______

Coefficient of reproducibility = 0.99

Coefficient of scalability = 0.98

Provided by the Internet Stroke Center — www.strokecenter.org

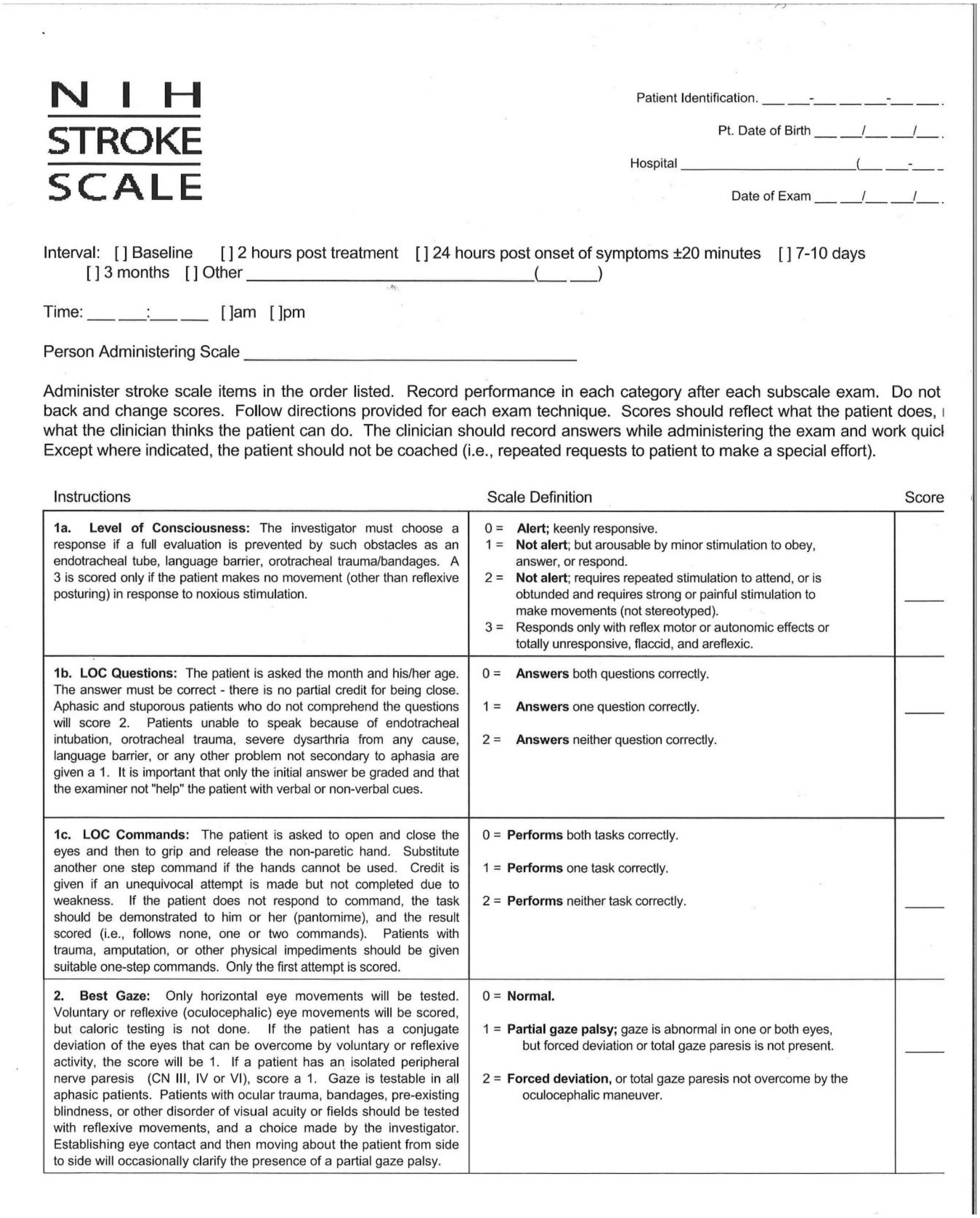


Grossmt (Gross Movement)

1. Place hand behind head (If score = 3, total = 9 and finish) ________
2. (If score = 0, total = 0 and finish ________
3. Place hand on top of head ________
4. Hand to mouth ________

Coefficient of reproducibility = 0.98

Coefficient of scalability = 0.97



N I H STROKE SCALE

Patient Identification. ___ ___-___ ___ ___-___ ___.

Pt. Date of Birth ___ ___/___ ___/___.

Hospital ______________________(___ ___-___ _

Date of Exam ___ ___/___ ___/___.

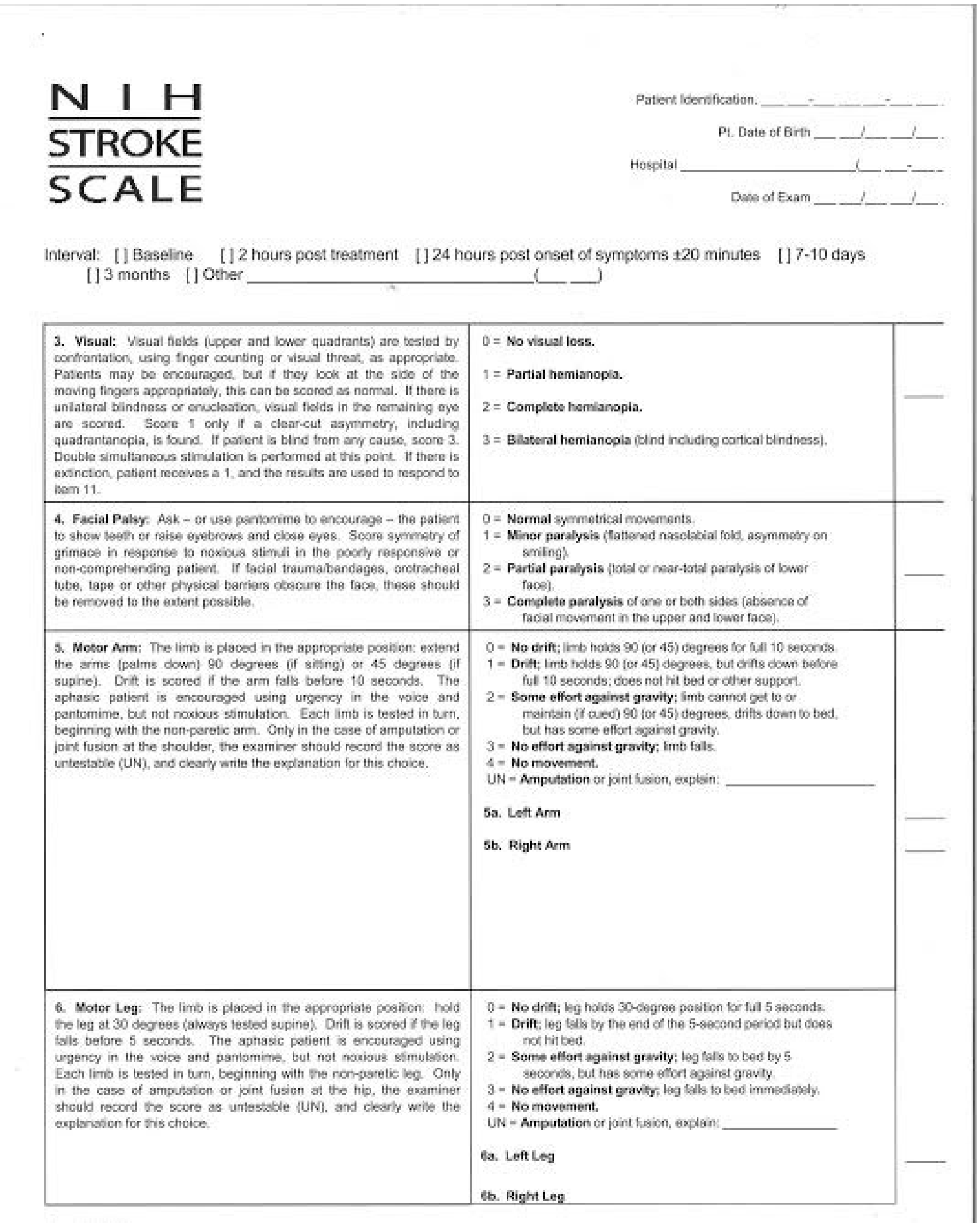
Interval: [ ] Baseline [ ] 2 hours post treatment [ ] 24 hours post onset of symptoms ±20 minutes [ ] 7-10 days
[ ] 3 months [ ] Other ______________________________(___ ___)

Time: ___ ___:___ ___ [ ]am [ ]pm

Person Administering Scale ______________________________

Administer stroke scale items in the order listed. Record performance in each category after each subscale exam. Do not back and change scores. Follow directions provided for each exam technique. Scores should reflect what the patient does, ı what the clinician thinks the patient can do. The clinician should record answers while administering the exam and work quicl Except where indicated, the patient should not be coached (i.e., repeated requests to patient to make a special effort).

Instructions	Scale Definition	Score
1a. Level of Consciousness: The investigator must choose a response if a full evaluation is prevented by such obstacles as an endotracheal tube, language barrier, orotracheal trauma/bandages. A 3 is scored only if the patient makes no movement (other than reflexive posturing) in response to noxious stimulation.	0 = **Alert;** keenly responsive. 1 = **Not alert**; but arousable by minor stimulation to obey, answer, or respond. 2 = **Not alert**; requires repeated stimulation to attend, or is obtunded and requires strong or painful stimulation to make movements (not stereotyped). 3 = Responds only with reflex motor or autonomic effects or totally unresponsive, flaccid, and areflexic.	____
1b. LOC Questions: The patient is asked the month and his/her age. The answer must be correct - there is no partial credit for being close. Aphasic and stuporous patients who do not comprehend the questions will score 2. Patients unable to speak because of endotracheal intubation, orotracheal trauma, severe dysarthria from any cause, language barrier, or any other problem not secondary to aphasia are given a 1. It is important that only the initial answer be graded and that the examiner not "help" the patient with verbal or non-verbal cues.	0 = **Answers** both questions correctly. 1 = **Answers** one question correctly. 2 = **Answers** neither question correctly.	____
1c. LOC Commands: The patient is asked to open and close the eyes and then to grip and release the non-paretic hand. Substitute another one step command if the hands cannot be used. Credit is given if an unequivocal attempt is made but not completed due to weakness. If the patient does not respond to command, the task should be demonstrated to him or her (pantomime), and the result scored (i.e., follows none, one or two commands). Patients with trauma, amputation, or other physical impediments should be given suitable one-step commands. Only the first attempt is scored.	0 = **Performs** both tasks correctly. 1 = **Performs** one task correctly. 2 = **Performs** neither task correctly.	____
2. Best Gaze: Only horizontal eye movements will be tested. Voluntary or reflexive (oculocephalic) eye movements will be scored, but caloric testing is not done. If the patient has a conjugate deviation of the eyes that can be overcome by voluntary or reflexive activity, the score will be 1. If a patient has an isolated peripheral nerve paresis (CN III, IV or VI), score a 1. Gaze is testable in all aphasic patients. Patients with ocular trauma, bandages, pre-existing blindness, or other disorder of visual acuity or fields should be tested with reflexive movements, and a choice made by the investigator. Establishing eye contact and then moving about the patient from side to side will occasionally clarify the presence of a partial gaze palsy.	0 = **Normal.** 1 = **Partial gaze palsy;** gaze is abnormal in one or both eyes, but forced deviation or total gaze paresis is not present. 2 = **Forced deviation,** or total gaze paresis not overcome by the oculocephalic maneuver.	____



NIH STROKE SCALE

Patient Identification. ___ ___-___ ___ ___-___ ___ .

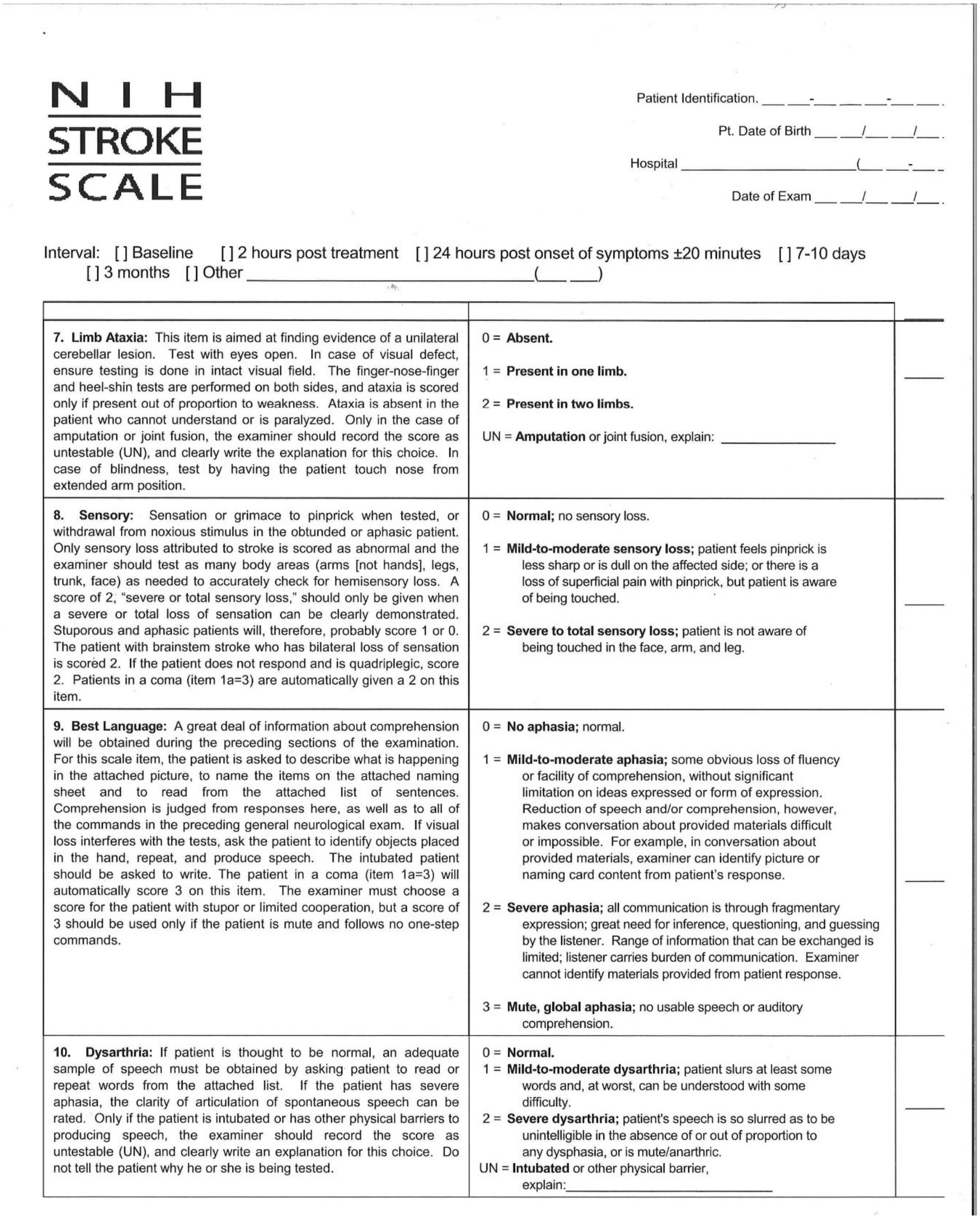
Pt. Date of Birth ___ ___/___ ___/___ .

Hospital ______________________(___ ___-___ .

Date of Exam ___ ___/___ ___/___ .

Interval: [ ] Baseline [ ] 2 hours post treatment [ ] 24 hours post onset of symptoms ±20 minutes [ ] 7-10 days
[ ] 3 months [ ] Other ______________________________(___ ___)

3. Visual: Visual fields (upper and lower quadrants) are tested by confrontation, using finger counting or visual threat, as appropriate. Patients may be encouraged, but if they look at the side of the moving fingers appropriately, this can be scored as normal. If there is unilateral blindness or enucleation, visual fields in the remaining eye are scored. Score 1 only if a clear-cut asymmetry, including quadrantanopia, is found. If patient is blind from any cause, score 3. Double simultaneous stimulation is performed at this point. If there is extinction, patient receives a 1, and the results are used to respond to item 11.	0 = **No visual loss.** 1 = **Partial hemianopia.** 2 = **Complete hemianopia.** 3 = **Bilateral hemianopia** (blind including cortical blindness).	____
4. Facial Palsy: Ask – or use pantomime to encourage – the patient to show teeth or raise eyebrows and close eyes. Score symmetry of grimace in response to noxious stimuli in the poorly responsive or non-comprehending patient. If facial trauma/bandages, orotracheal tube, tape or other physical barriers obscure the face, these should be removed to the extent possible.	0 = **Normal** symmetrical movements. 1 = **Minor paralysis** (flattened nasolabial fold, asymmetry on smiling). 2 = **Partial paralysis** (total or near-total paralysis of lower face). 3 = **Complete paralysis** of one or both sides (absence of facial movement in the upper and lower face).	____
5. Motor Arm: The limb is placed in the appropriate position: extend the arms (palms down) 90 degrees (if sitting) or 45 degrees (if supine). Drift is scored if the arm falls before 10 seconds. The aphasic patient is encouraged using urgency in the voice and pantomime, but not noxious stimulation. Each limb is tested in turn, beginning with the non-paretic arm. Only in the case of amputation or joint fusion at the shoulder, the examiner should record the score as untestable (UN), and clearly write the explanation for this choice.	0 = **No drift;** limb holds 90 (or 45) degrees for full 10 seconds. 1 = **Drift;** limb holds 90 (or 45) degrees, but drifts down before full 10 seconds; does not hit bed or other support. 2 = **Some effort against gravity;** limb cannot get to or maintain (if cued) 90 (or 45) degrees, drifts down to bed, but has some effort against gravity. 3 = **No effort against gravity;** limb falls. 4 = **No movement.** UN = **Amputation** or joint fusion, explain: ____________________ **5a. Left Arm** **5b. Right Arm**	____ ____
6. Motor Leg: The limb is placed in the appropriate position: hold the leg at 30 degrees (always tested supine). Drift is scored if the leg falls before 5 seconds. The aphasic patient is encouraged using urgency in the voice and pantomime, but not noxious stimulation. Each limb is tested in turn, beginning with the non-paretic leg. Only in the case of amputation or joint fusion at the hip, the examiner should record the score as untestable (UN), and clearly write the explanation for this choice.	0 = **No drift;** leg holds 30-degree position for full 5 seconds. 1 = **Drift;** leg falls by the end of the 5-second period but does not hit bed. 2 = **Some effort against gravity;** leg falls to bed by 5 seconds, but has some effort against gravity. 3 = **No effort against gravity;** leg falls to bed immediately. 4 = **No movement.** UN = **Amputation** or joint fusion, explain: ________________ **6a. Left Leg** **6b. Right Leg**	____



NIH STROKE SCALE

Patient Identification. __ __-__ __ __-__ __.

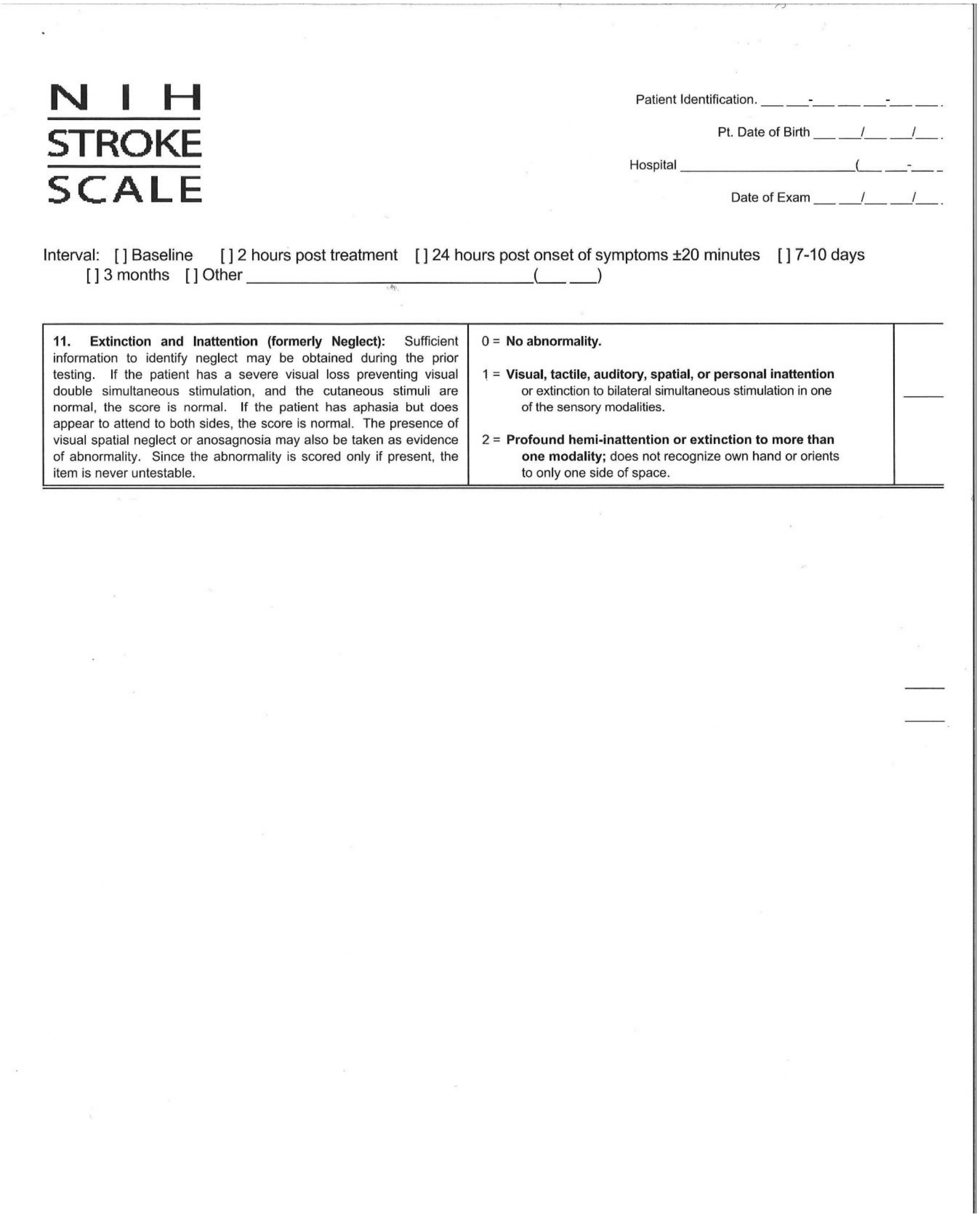
Pt. Date of Birth __ __/__ __/__.

Hospital ________________(__ __-__ _

Date of Exam __ __/__ __/__.

Interval: [ ] Baseline [ ] 2 hours post treatment [ ] 24 hours post onset of symptoms ±20 minutes [ ] 7-10 days
[ ] 3 months [ ] Other ____________________(__ __)

7. Limb Ataxia: This item is aimed at finding evidence of a unilateral cerebellar lesion. Test with eyes open. In case of visual defect, ensure testing is done in intact visual field. The finger-nose-finger and heel-shin tests are performed on both sides, and ataxia is scored only if present out of proportion to weakness. Ataxia is absent in the patient who cannot understand or is paralyzed. Only in the case of amputation or joint fusion, the examiner should record the score as untestable (UN), and clearly write the explanation for this choice. In case of blindness, test by having the patient touch nose from extended arm position.	0 = **Absent.** 1 = **Present in one limb.** 2 = **Present in two limbs.** UN = **Amputation** or joint fusion, explain: __________	____
8. Sensory: Sensation or grimace to pinprick when tested, or withdrawal from noxious stimulus in the obtunded or aphasic patient. Only sensory loss attributed to stroke is scored as abnormal and the examiner should test as many body areas (arms [not hands], legs, trunk, face) as needed to accurately check for hemisensory loss. A score of 2, "severe or total sensory loss," should only be given when a severe or total loss of sensation can be clearly demonstrated. Stuporous and aphasic patients will, therefore, probably score 1 or 0. The patient with brainstem stroke who has bilateral loss of sensation is scored 2. If the patient does not respond and is quadriplegic, score 2. Patients in a coma (item 1a=3) are automatically given a 2 on this item.	0 = **Normal;** no sensory loss. 1 = **Mild-to-moderate sensory loss;** patient feels pinprick is less sharp or is dull on the affected side; or there is a loss of superficial pain with pinprick, but patient is aware of being touched. 2 = **Severe to total sensory loss;** patient is not aware of being touched in the face, arm, and leg.	____
9. Best Language: A great deal of information about comprehension will be obtained during the preceding sections of the examination. For this scale item, the patient is asked to describe what is happening in the attached picture, to name the items on the attached naming sheet and to read from the attached list of sentences. Comprehension is judged from responses here, as well as to all of the commands in the preceding general neurological exam. If visual loss interferes with the tests, ask the patient to identify objects placed in the hand, repeat, and produce speech. The intubated patient should be asked to write. The patient in a coma (item 1a=3) will automatically score 3 on this item. The examiner must choose a score for the patient with stupor or limited cooperation, but a score of 3 should be used only if the patient is mute and follows no one-step commands.	0 = **No aphasia;** normal. 1 = **Mild-to-moderate aphasia;** some obvious loss of fluency or facility of comprehension, without significant limitation on ideas expressed or form of expression. Reduction of speech and/or comprehension, however, makes conversation about provided materials difficult or impossible. For example, in conversation about provided materials, examiner can identify picture or naming card content from patient's response. 2 = **Severe aphasia;** all communication is through fragmentary expression; great need for inference, questioning, and guessing by the listener. Range of information that can be exchanged is limited; listener carries burden of communication. Examiner cannot identify materials provided from patient response. 3 = **Mute, global aphasia;** no usable speech or auditory comprehension.	____
10. Dysarthria: If patient is thought to be normal, an adequate sample of speech must be obtained by asking patient to read or repeat words from the attached list. If the patient has severe aphasia, the clarity of articulation of spontaneous speech can be rated. Only if the patient is intubated or has other physical barriers to producing speech, the examiner should record the score as untestable (UN), and clearly write an explanation for this choice. Do not tell the patient why he or she is being tested.	0 = **Normal.** 1 = **Mild-to-moderate dysarthria;** patient slurs at least some words and, at worst, can be understood with some difficulty. 2 = **Severe dysarthria;** patient's speech is so slurred as to be unintelligible in the absence of or out of proportion to any dysphasia, or is mute/anarthric. UN = **Intubated** or other physical barrier, explain:______________	____



NIH STROKE SCALE

Patient Identification. ___ ___-___ ___ ___-___ ___.

Pt. Date of Birth ___ ___/___ ___/___.

Hospital ______________________(___ ___-___ _

Date of Exam ___ ___/___ ___/___.

Interval: [ ] Baseline [ ] 2 hours post treatment [ ] 24 hours post onset of symptoms ±20 minutes [ ] 7-10 days
[ ] 3 months [ ] Other ______________________________(___ ___)

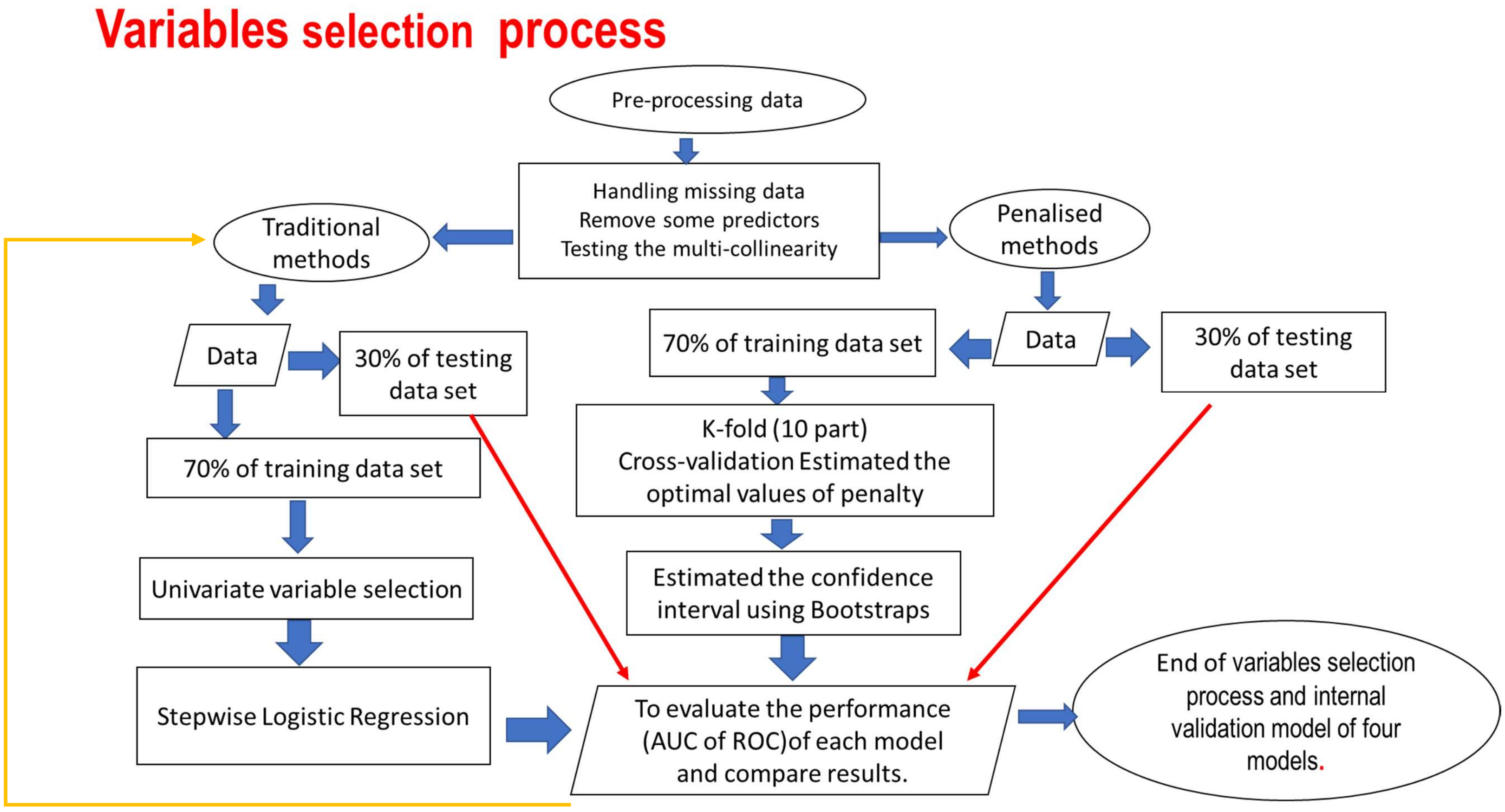
11. Extinction and Inattention (formerly Neglect): Sufficient information to identify neglect may be obtained during the prior testing. If the patient has a severe visual loss preventing visual double simultaneous stimulation, and the cutaneous stimuli are normal, the score is normal. If the patient has aphasia but does appear to attend to both sides, the score is normal. The presence of visual spatial neglect or anosagnosia may also be taken as evidence of abnormality. Since the abnormality is scored only if present, the item is never untestable.	0 = **No abnormality.** 1 = **Visual, tactile, auditory, spatial, or personal inattention** or extinction to bilateral simultaneous stimulation in one of the sensory modalities. 2 = **Profound hemi-inattention or extinction to more than one modality;** does not recognize own hand or orients to only one side of space.	_____

_____

_____

10.3 Appendix C

10.3.1 Flow chart of variable selection algorithm



Variables selection process

Pre-processing data

Handling missing data
Remove some predictors
Testing the multi-collinearity

Traditional methods

Data

30% of testing data set

70% of training data set

Univariate variable selection

Stepwise Logistic Regression

Penalised methods

Data

70% of training data set

30% of testing data set

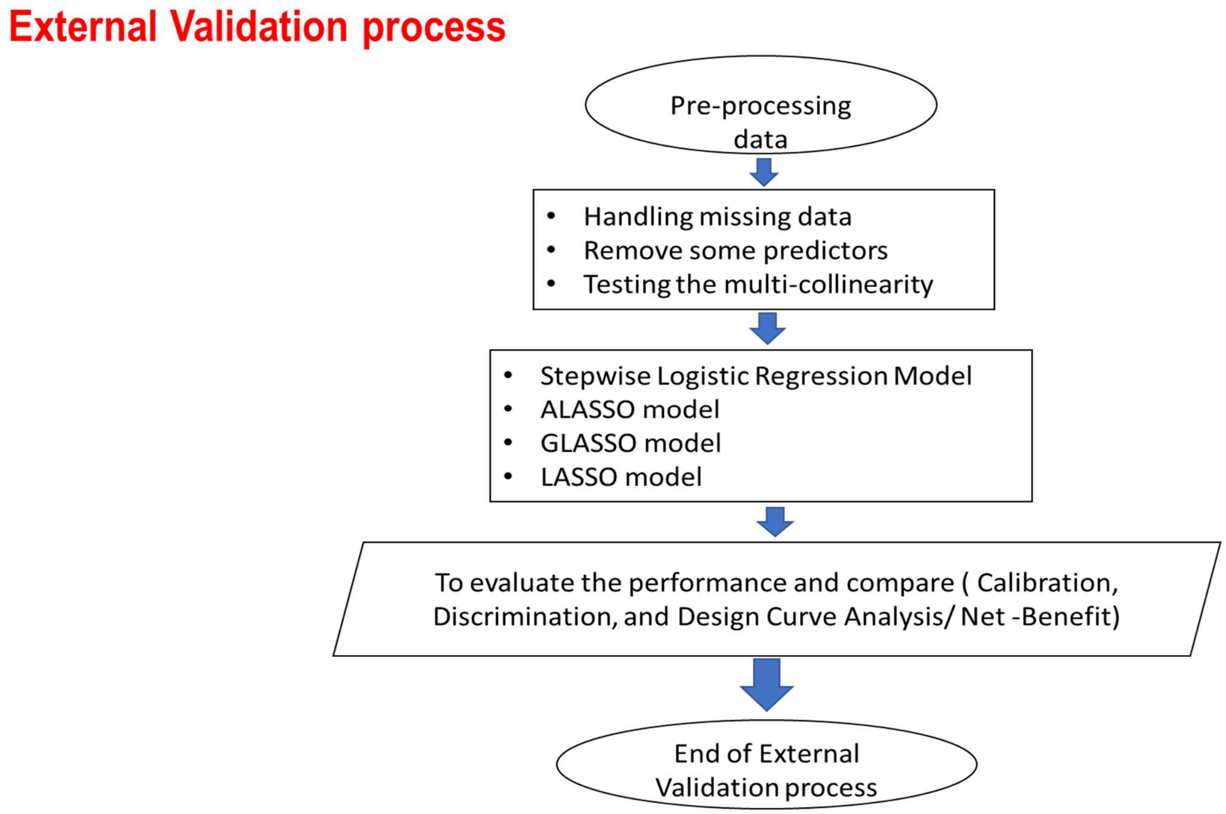
K-fold (10 part)
Cross-validation Estimated the optimal values of penalty

Estimated the confidence interval using Bootstraps

To evaluate the performance (AUC of ROC)of each model and compare results.

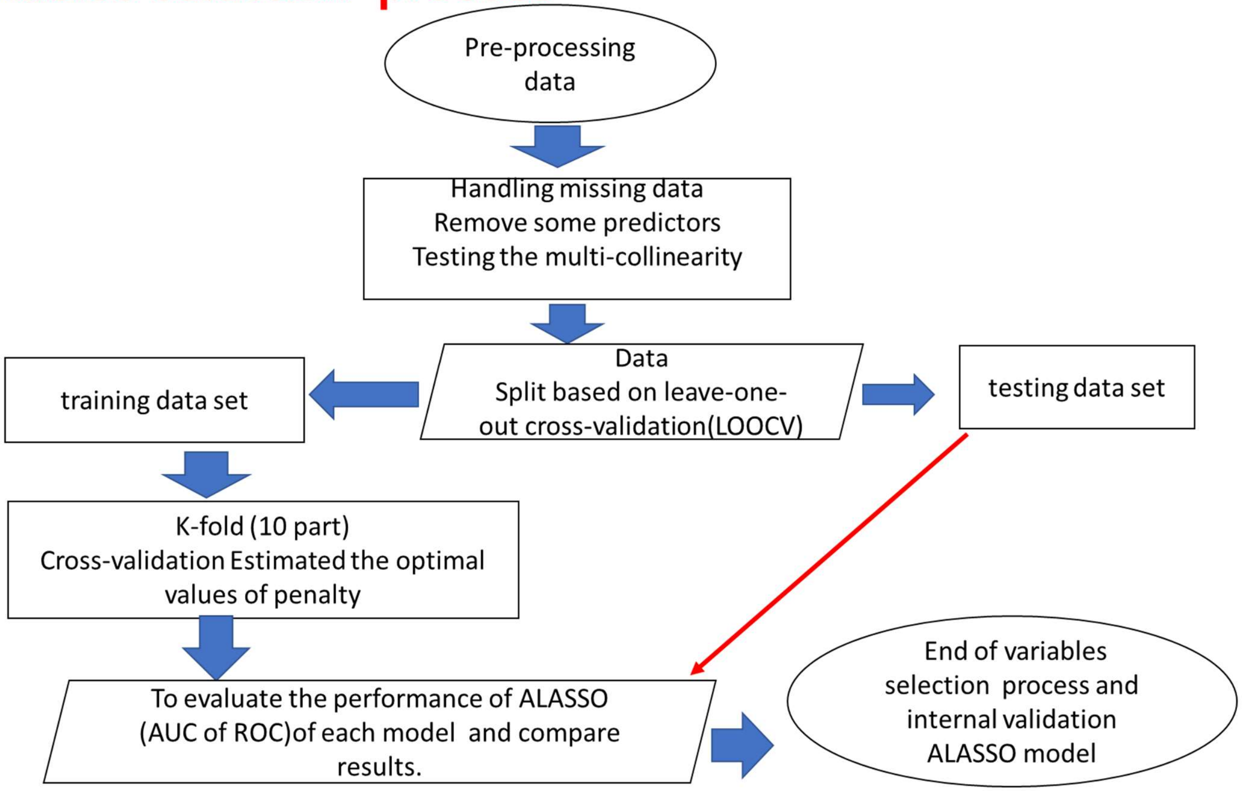
End of variables selection process and internal validation model of four models.

10.3.2 Flow chart of external validation algorithm


External Validation process
Pre-processing data
• Handling missing data
• Remove some predictors
• Testing the multi-collinearity
• Stepwise Logistic Regression Model
• ALASSO model
• GLASSO model
• LASSO model
To evaluate the performance and compare (Calibration, Discrimination, and Design Curve Analysis/ Net -Benefit)
End of External Validation process

10.3.3 Flow chart of variable selection algorithm

Variables selection process



Pre-processing data

Handling missing data
Remove some predictors
Testing the multi-collinearity

Data
Split based on leave-one-out cross-validation(LOOCV)

training data set

testing data set

K-fold (10 part)
Cross-validation Estimated the optimal values of penalty

To evaluate the performance of ALASSO (AUC of ROC)of each model and compare results.

End of variables selection process and internal validation ALASSO model